ARI TECHNICAL REPORT
TR-79-A6

LEVEL

# Simulation of a Model Tank Gunnery Test

by

Paul W. Fingerman and George R. Wheaton

AMERICAN INSTITUTES FOR RESEARCH

and

G. Gary Boycan

ARI Engagement Simulation Technical Area





March 1979



Contract DAHC-19-76-C-0003

Prepared for

U.S. ARMY RESEARCH INSTITUTE
for the BEHAVIORAL and SOCIAL SCIENCES
5001 Eisenhower Avenue
Alexandria, Virginia 22333

# U. S. ARMY RESEARCH INSTITUTE
# FOR THE BEHAVIORAL AND SOCIAL SCIENCES

## A Field Operating Agency under the Jurisdiction of the Deputy Chief of Staff for Personnel

JOSEPH ZEIDNER
Technical Director

WILLIAM L. HAUSER
Colonel, U S Army
Commander

NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U. S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-P, 5001 Eisenhower Avenue, Alexandria, Virginia 22333.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U. S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

| REPORT DOCUMENTATION PAGE | READ INSTRUCTIONS BEFORE COMPLETING FORM |
|---|---|
| 1. REPORT NUMBER: TR-79-A6 (19) (18) ARI — 2. GOVT ACCESSION NO. | 3. RECIPIENT'S CATALOG NUMBER |
| 4. TITLE (and Subtitle): (6) SIMULATION OF A MODEL TANK GUNNERY TEST. | 5. TYPE OF REPORT & PERIOD COVERED: Final Report 9/77 - 9/78 |
| | 6. PERFORMING ORG. REPORT NUMBER: (14) AIR-55800-9/78-FR |
| 7. AUTHOR(s): (10) Paul W. Fingerman, George R. Wheaton, G. Gary Boycan, ~~& ARI~~ | 8. CONTRACT OR GRANT NUMBER(s): (15) DAHC 19-76-C-0003 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS: American Institutes for Research, 1055 Thomas Jefferson St., NW, Washington, DC 20007 — 023 450 | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: (16) 2Q163743A780 |
| 11. CONTROLLING OFFICE NAME AND ADDRESS: US Army Research Institute for the Behavioral and Social Sciences, 5001 Eisenhower Avenue, Alexandria, VA 22333 (11) | 12. REPORT DATE: March 1979 — 13. NUMBER OF PAGES: 74 |
| 14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): (12) 84 p. | 15. SECURITY CLASS. (of this report): Unclassified — 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE |

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

(9) Final rept. Sep 77-Sep 78,

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Criterion-referenced testing
simulation
tank gunnery
crew marksmanship
evaluation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

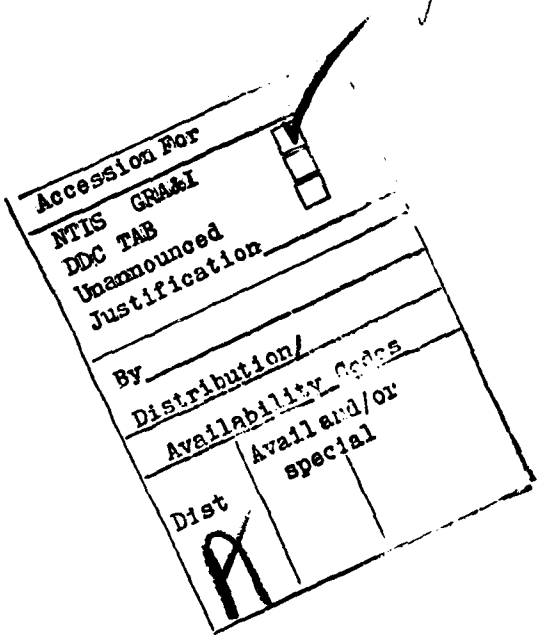
This report describes the activities conducted during Phase II of a project devoted to the development of methods for assessing tank crew marksmanship performance. An earlier report on Phase I (ARI report TR 78-A24, AD A061 153) presented methods for the development of a criterion-referenced test of marksmanship. This effort led to the specification of a model livefire test of tank crew gunnery, together with scoring and test administration procedures for determining crew qualification. The present effort was aimed at evaluating

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified 023 450

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. the feasibility of assessing crew gunnery through the use of simulated, rather than livefired, test exercises. Thirty-nine simulator devices were evaluated; two were identified as viable candidates for use in a simulated test. The analytic methodology is described, and a proposed simulated model test is presented. The report concludes with a discussion of the evaluation procedures that are required if the simulated model test is to be considered for use as a substitute for the livefire test.



## ACKNOWLEDGMENTS

The authors gratefully acknowledge the contributions of the many individuals who assisted in the research reported herein. Lieutenant Colonel R. M. Bosserman and Mr. Robert K. Bauer of the Directorate of Training Developments, U.S. Army Armor School at Fort Knox aided greatly in the conceptualization of Armor training and testing throughout the project. Staff Sergeant Rodney Caesar, of the Main Battle Tank Branch, Weapons Department, Armor School, deserves our special thanks for his painstaking and insightful review of the candidate simulators and their capabilities. His special efforts were critical to the success of the project.

# SIMULATION OF A MODEL TANK GUNNERY TEST

## BRIEF

Requirement:

To develop a simulated version of a model livefire tank gunnery test that can be used to evaluate crew marksmanship. By using simulation to reduce testing costs, more engagements can be added to increase confidence in the accuracy of crew qualification decisions and to provide more diagnostic information about crew training needs.

Procedure:

The feasibility of using simulation techniques as cost-effective alternatives to livefire testing was examined for a model livefire test of tank gunnery developed in Phase 1 of the project. The model test takes into consideration different types of target engagements as well as the behaviors of the individual crew members that are required during firing. The goal was to identify existing simulators or training devices that might best be used to simulate some or all of the model livefire test exercises.

The analysis consisted of several steps. Devices and simulators were identified from a variety of sources and compiled into an initial list of 39 candidates. These were screened to cull out those failing to meet certain basic requirements. The 14 remaining candidates were then evaluated with respect to the specific types of engagement conditions they could simulate. This information was then used to examine three finalist devices with respect to the specific exercises that they could simulate, the exercises being drawn from the model test and various versions of Table VIII. The quality of simulation was determined by considering the crew behaviors in each exercise and assessing the extent to which each behavior would be produced in the simulated version. A simulated test was then constructed that appeared promising enough to warrant empirical tryout and evaluation.

Test evaluation issues were also studied. Development of a formal plan for empirically assessing the reliability and validity of the simulated marksmanship test was viewed as absolutely essential.

Findings:

Two devices were identified that potentially could be used to simulate a test of crew marksmanship. One device consists of the M2 .50 caliber machinegun affixed to an operational M60A1AØS tank by means of a Telfare mount. It can be used on a 1/2-scale or full-scale range facility. The other device is the Tank Appended Crew Evaluation Device (TACED) which provides gun camera pictures of the crew's aiming performance. Because of their complementary features, it is recommended that both be used to support the simulated testing of livefire main gun engagements.

The simulated test is a variant of the model livefire test. There are 13 simulated main gun exercises and 15 live-fired machinegun exercises on which crews are tested during daylight and nighttime conditions. However, each of the simulated main gun engagements is repeated three times in an attempt to increase test reliability. The resulting 54 test exercises are designed to provide estimates of crew proficiency on the 266 objectives comprising the domain of tank gunnery.

The accompanying evaluation plan provides for an empirical assessment of the test using one of two different experimental designs. The pros and cons of each design are discussed as are practical considerations such as the requirements for the number of crews participating, and the advisability of different shortcuts when implementing the evaluation study.

Utilization of Findings:

Crew proficiency in the use of tank weapons is a major goal of gunnery training and evaluation. Recently, however, the Army has decided to shift much of the ammunition allocated for training and evaluation from crew marksmanship to section, platoon, and higher element tactical gunnery. The simulated model test of crew marksmanship can expedite this shift while improving evaluation of crew proficiency in operating the tank weapon system. To reach this objective the empirical field study recommended in the report should be carried out.

TABLE OF CONTENTS

TABLE OF CONTENTS (cont'd.)

LIST OF TABLES

# I. INTRODUCTION

The U.S. Army must maintain and document the readiness of its armored forces to move, shoot, and coordinate at every tactical level. The foundation for this readiness is the capacity of the tank crew to use all of its weaponry to neutralize a variety of threat targets.

In order to evaluate their proficiency at neutralizing targets, the U.S. Army requires its tank crews to qualify annually in gunnery. The specific gunnery exercises that have been used to evaluate crew proficiency, together with the scoring systems and qualification standards that have been applied to their performance, are defined by Table VIII in FM 17-12 (1977) and FM 17-12-2 (1977). These exercises have been selected and developed on the basis of competent opinion and the judgment of experienced armor personnel who, believing that comprehensive and exhaustive testing of crew capabilities is impossible because of resource constraints, have attempted to distill the essence of gunnery into a manageable set of test exercises. But, test developers have included and excluded exercises from past, present, and proposed gunnery tables without making explicit the rationales for doing so. Moreover, they have relied on the costly use of full-caliber main gun ammunition, where simulation techniques might have been equally effective. In the past these costs have ranged from $101.00 for a HEP-TP-T round to $242.00 for a SABOT service round. The total cost of the 27 main gun rounds allocated for one firing of the Table VIII described in FM 17-12-2 (1977) is about $4000; approximately $200,000 per battalion.

In response to these problems the U.S. Army Research Institute under U.S. Army Armor School sponsorship has supported a program of research concerned with the development of cost-effective techniques for evaluating crew weapons proficiency. The need for such research was verified during Phase I of the project (Wheaton, Fingerman, and Boycan, 1978). A review of gunnery tables indicated that there were a number of redundant exercises on the one hand, while on the other, entire classes of exercises had been omitted from the tests. Nor were the exercises representative of the entire set of crew gunnery behaviors. Therefore, one could have little confidence when making general inferences about a crew's qualifications based on its (limited and unrepresentative) test performance.

To address this problem a set of systematic, analytic techniques was developed for constructing crew qualification tests. The techniques were used to recommend exercises for

inclusion in a model test of crew marksmanship. Emphasis was given to the development of explicit testing rationales, together with associated scoring procedures, based on the issues of test content and purpose. The objective agreed upon was construction of a test that would be optimal for crew qualification but that would also serve a variety of other purposes.

The 28-item test that was ultimately developed satisfied this objective and met a number of other criteria considered essential to the design of an effective test. First, at least one highly representative exercise was included from each major family of gunnery objectives. This step provided a basis for inferences to be drawn about the quality of performance in each family and about proficiency in the gunnery domain as a whole. Second, the exercises spanned the range of firing conditions under which engagements might occur. Third, the test exercises required the crew to demonstrate their ability to perform nearly all of the 112 crew behaviors involved in tank gunnery. Most importantly, the resulting marksmanship table served as a model example of the kind of highly effective test that could be developed in other settings with other weapon systems, were the same explicit test construction techniques used.

## PURPOSE OF THE RESEARCH

The research described in this report represents the second phase of the larger program mentioned above. The second phase of the effort examined the feasibility of using simulation techniques as cost-effective alternatives to live-fire testing. Impetus for the examination came from the Army's recent decision to invest more of its ammunition allocation in the evaluation of section, platoon, and higher echelon tactical gunnery. If feasible, the use of simulation techniques in assessing crew marksmanship would clearly expedite such a shift. At the same time, having drastically reduced crew qualification testing costs, more engagements could be added to the crew marksmanship table to increase confidence in the accuracy of qualification decisions, and to provide increased diagnostic information about crew strengths and weaknesses.

The remainder of this report describes the search for viable simulations of a marksmanship table and discusses the steps required to evaluate their effectiveness as tests. In the next section candidate devices are identified and the process and results of a rigorous screening program are described. The potential of the most useful devices for design of simulated marksmanship tests is then discussed. The last major section deals with the complex issue of device evaluation.

Traditional concepts of reliability and validity are discussed in light of the nontraditional criterion-referenced marksmanship table. Experimental designs and resource requirements for evaluation procedures are then described. This discussion is intended to be of direct benefit to those who must empirically evaluate the feasibility of replacing livefire with simulation alternatives.

## II. ANALYSIS OF GUNNERY SIMULATORS

Given a model livefire test of crew marksmanship (Tables 1 and 2), the focus of the project shifted to the issue of cost-effectiveness in marksmanship testing. Specifically, the effectiveness of using simulation techniques in place of expensive livefire testing was examined.

The goal was to identify existing simulators or training devices that might be used to simulate some or all of the model livefire exercises, and to develop a simulated version of the test that could be substituted for livefire exercises, especially those involving the main gun. To accomplish this objective an analysis was undertaken consisting of six steps: Devices and simulators were identified and compiled into an initial list of candidates. These were screened to cull out those failing to meet certain basic requirements. The remaining devices were then evaluated with respect to the specific types of engagement conditions that they could or could not simulate. This information was used to evaluate each of the still viable candidates vis à vis each specific exercise in the model test, as well as exercises included in various versions of Table VIII. The quality with which each exercise was simulated was determined by considering the crew behaviors involved in each exercise, and assessing the extent to which each behavior would be reproduced in the simulated version. The final step was to construct simulated model tests that appeared promising enough to warrant empirical tryout and evaluation. Each of the six steps is described in the remainder of this section.

### STEP 1 - COMPILATION OF DEVICES

Candidate training devices and simulators were gleaned from several sources. Many types of relevant documentation were surveyed including Chapter 18 of FM 17-12 (1977), Draft TC 17-12-7 (1976), TRADOC Pam 71-9 (1976), the Index and Description of Army Training Devices (DOA Pam 310-12, 1972), and relevant articles in Armor magazine. These materials were then reviewed and augmented during discussions with cognizant subject matter experts from the Devices Branch, Collective Training Division, Directorate of Training Developments (DTD), and the Weapons Department, U.S. Army Armor School (USAARMS), Fort Knox, Kentucky. The search resulted in the identification of some 39 devices that held varying initial degrees of promise for implementation of simulated marksmanship exercises.

**Table 1. Model Test of Crew Marksmanship: Part I. Main Gun**

Unit — Tank Crew — TC

Gunner — Loader — Driver

| | | EXERCISE | CONDITIONS | NO. OF ROUNDS | AMMUNITION | STANDARD |
|---|---|---|---|---|---|---|
| Day Time | 1 | MOVING TANK | Battlesight<br>1600 m<br>Moving to a halt<br>Telescope | 2 | TPDS-T | Engage in 5 seconds.<br>Hit in 10 seconds. |
| | 2 | TANK FRONT SHOT | Battlesight<br>1000m<br>On the move<br>Gunner's periscope | 2 | HEAT-TP-T | Engage in 5 seconds.<br>Hit in 10 seconds. |
| | 3 | MOVING TANK (TC) | Precision<br>1700 m<br>Moving to a halt<br>Rangefinder | 2 | TPDS-T | Engage in 10 seconds.<br>Hit in 15 seconds. |
| | 4 | MOVING TANK | Precision<br>1700 m<br>Moving to a halt<br>Telescope | 2 | HEAT-TP-T | Engage in 10 seconds.<br>Hit in 15 seconds. |
| | 5 | TANK SILHOUETTE | Precision<br>2000 m<br>Moving to a halt<br>Gunner's periscope | 2 | TPDS-T | Engage in 10 seconds.<br>Hit in 15 seconds. |
| | 6 | BUNKER/ CREW WEAPON (TC) | Precision<br>2200 m<br>Moving to a halt<br>Rangefinder | 2 | HEP-TP-T | Engage in 10 seconds.<br>Hit in 15 seconds. |
| | 7 | MOVING TRUCK | Precision<br>1200 m<br>Moving to a halt<br>Telescope | 2 | HEP-TP-T | Engage in 10 seconds.<br>Hit in 15 seconds. |
| Night Time | 1 | MOVING TANK | Range card lay to direct fire<br>1900 m<br>Stationary vehicle<br>Telescope, flare | 2 | HEAT-TP-T | Hit within 5 minutes of reaching referenced position. |
| | 2 | TROOPS | Range card lay to direct fire<br>900 m<br>Stationary vehicle<br>Gunner Periscope, infrared | 2 | APERS | Hit within 5 minutes of reaching referenced position. |
| | 3 | MOVING TANK (TC) | Range card lay to direct fire<br>1400 m<br>Stationary vehicle<br>Rangefinder, flare | 2 | TPDS-T | Hit within 5 minutes of reaching referenced position. |
| | 4 | TANK FRONT SHOT | Battlesight<br>800 m<br>Stationary vehicle<br>Gunner's Periscope, infrared | 2 | HEAT-TP-T | Engage in 10 seconds.<br>Hit within 15 seconds. |
| | 5 | TROOPS | Precision<br>1700 m<br>Moving to a halt<br>Gunner's periscope, flare | 2 | APERS | Engage in 15 seconds.<br>Hit within 20 seconds. |
| | 6 | TANK FRONT SHOT (TC) | Battlesight<br>1300 m<br>On the move<br>Rangefinder, flare | 2 | TPDS-T | Engage in 10 seconds<br>Hit within 15 seconds. |

NOTES
1. During conduct of the table, target acquisition time, time to first-round hit, and time to second-round hit (if needed) are recorded. Scoring is then accomplished using a variety of procedures.
2. Emphasis is on achieving a target hit in the shortest possible time. Bonus points are given for ammunition conservation, and second round is not fired if the first round hits.
3. Crew duties are NOT scored
4. Three main gun rounds have been allocated for warm-up and zero confirmation (two rounds for day, one round for night). The least expensive round (HEP-TP-T) should be used for warm-up purposes and the highest muzzle velocity ammunition (TPDS-T) should be used for zeroing.
5. As an alternative, the second night engagement may be fired at a moving truck with HEP
6. As an alternative, the fifth engagement may be fired at a bunker with HEP and the telescope
7. "Engage in 5 seconds" refers to the time from the alert element of the initial fire command or laying of the main gun for direction (whichever occurs earlier) to the firing of the first round. A second round, if needed, must be fired within 5 seconds of a first round miss.
8. Flare illumination may be replaced with white light illumination from another tank.

**Table 2. Model Test of Crew Marksmanship: Part II. Machinegun**

| | | EXERCISE | CONDITIONS | NUMBER ROUNDS | STANDARD |
|---|---|---|---|---|---|
| Day Time | 1 | TROOPS | 300 m<br>On the move<br>Infinity sight | 100<br>Coax | Engage within 5 seconds.<br>Obtain 3/5s coverage. |
| | 2 | TRUCK | 900 m<br>Moving to a halt<br>Rangefinder | 50<br>Coax | Engage within 5 seconds.<br>Obtain 1 tracer hit. |
| | 3 | MOVING TRUCK | 700 m<br>On the move<br>Infinity sight | 50<br>Coax | Engage within 5 seconds.<br>Obtain 1 tracer hit. |
| | 4 | AIRCRAFT | 2200 m<br>Moving to a halt<br>Tank Commander's periscope | 100<br>Cal .50 | Engage within 5 seconds.<br>Obtain 1 tracer hit. |
| | 5 | TRUCK | 500 m<br>On the move<br>Infinity sight | 50<br>Coax | Engage within 5 seconds.<br>Obtain 1 tracer hit. |
| | 6 | TROOP CARRIER | 1500 m<br>On the move<br>Tank Commander's periscope | 100<br>Cal .50 | Engage within 5 seconds.<br>Obtain 3/5s coverage. |
| Night Time | 1 | MOVING TRUCK | 300 m<br>Stationary vehicle<br>Gunner's periscope, infrared, RCLD | 50<br>Coax | Engage within 10 seconds.<br>Obtain 1 tracer hit. |
| | 2 | TRUCK | 500 m<br>Stationary vehicle<br>Metascope, infrared, RCLD | 50<br>Coax | Engage within 10 seconds.<br>Obtain 1 tracer hit. |
| | 3 | TRUCK | 900 m<br>Moving to a halt<br>Infinity sight, flare | 50<br>Coax | Engage within 10 seconds.<br>Obtain 1 tracer hit. |
| | 4 | AIRCRAFT | 900 m<br>Moving to a halt<br>Tank Commander's periscope, infrared | 100<br>Cal .50 | Engage within 10 seconds.<br>Obtain 1 tracer hit. |
| | 5 | TROOPS | 700 m<br>On the move<br>Gunner's periscope, infrared | 100<br>Coax | Engage within 10 seconds.<br>Obtain 3/5s coverage. |
| | 6 | MOVING TRUCK | 300 m<br>Moving to a halt<br>Metascope, infrared | 50<br>Coax | Engage within 10 seconds.<br>Obtain 1 tracer hit. |
| | 7 | MOVING TRUCK | 500 m<br>Moving to a halt<br>Gunner's periscope, infrared | 50<br>Coax | Engage within 10 seconds.<br>Obtain 1 tracer hit. |
| | 8 | MOVING AIRCRAFT | 900 m<br>Moving to a halt<br>Tank Commander's periscope, infrared | 100<br>Cal .50 | Engage within 10 seconds.<br>Obtain 1 tracer hit. |
| | 9 | AIRCRAFT | 2000 m<br>Stationary vehicle<br>Tank Commander's periscope, flare | 100<br>Cal .50 | Engage within 10 seconds.<br>Obtain 1 tracer hit. |

NOTES: 1. During conduct of the table, engagement and hit times are recorded. Scoring is then accomplished using a variety of procedures.
2. As an alternative to Exercise 8, a light-armored vehicle may be engaged.
3. Flare illumination may be replaced with white light illumination from another tank. Daylight standards should then be used.
4. Metascope engagements are fired by the tank commander.

To facilitate processing in subsequent stages of the analysis, the candidates were assigned to one of three categories, each of which contained a relatively homogeneous set of devices. In the first category were 25 devices, each consisting of a specific kind of subcaliber weapon, a particular type of mount used to affix the weapon to an operational M60A1AØS tank, and a scaled-range facility. In any specific instance the latter was either a 1/60, 1/35, 1/20, or 1/2-scale range equipped with pop-up and knock-down scaled targets, or in the case of laser ranges, special retroreflective nonknock-down targets. The devices comprising this first category, and representing unique mixes of weapon, mount, and scaled range, are indicated in Table 3 by check marks. Empty cells correspond to cases clearly ruled out as viable candidates, either because target effects would be too severe (e.g., for some of the larger caliber weapons fired on the smaller ranges) or because the limits of accurate fire would be exceeded (e.g., for smaller caliber weapons used on the larger ranges). (As will be discussed shortly, cells with double check marks represent devices deemed potentially viable in a subsequent stage of the analysis.)

None of the devices in the second category involve actual livefiring, although roughly half do require range facilities of some kind (i.e., REALTRAIN, MILES, TACED, Stout, Dry Fire). The nine candidates represent a mixed bag of approaches to simulation as suggested by the listing in Table 4. Some, such as the venerable Green Hornet, have been in the inventory for years. Others, such as MILES, are still undergoing development.

The third and final category of devices consisted of five entries: the Unit Conduct-of-Fire Trainer, the Full Crew Interaction Simulator, the Tank Weapons Gunnery Simulation System, the Combat Training Theater (Subcaliber) and the Combat Training Theater (Laser). These candidates differ from those in the first two categories by virtue of the fact that they are still in early conceptual or engineering stages of development. Their implementation on even the most limited basis is not contemplated until some time in the 1980s. Thus, devices in this last set represent possibilities for the future.

An unexpected outcome of the search for simulators was the large number of potentially relevant devices that were uncovered. Most have been used as training devices for applications ranging from classroom instruction in the fundamentals of applying adjustment-of-fire procedures (e.g., the various conduct-of-fire trainers) to practice of combined-arms tactics in the field (e.g., REALTRAIN). Some, in addition to providing opportunities for practice of skills, have also been used to

**Table 3.**

**Candidate Scaled-Range Devices**

| Weapons and Mounts | Scaled Ranges 1/60 | 1/35 | 1/20 | 1/2 |
|---|---|---|---|---|
| BOI DVC-D 17-53 .22 Cal. In-Bore (90, 105mm) | √ | √ | | |
| DVC-D 17-85 .22 Cal. Mini-tank Bracket/Rifle | √ | √ | | |
| DVC-D 17-87 Brewster Device with M16 Rifle and .22 Cal. Rimfire Adapter | √√ | √√ | | |
| DVC-D 17-87 Brewster Device with M16 Rifle | √√ | √√ | √√ | √ |
| DVC-D 17-87 Brewster Device with 7.62mm Coaxial Machinegun | | | √ | √ |
| DVC-D 17-87 Brewster Device with M55 Laser Trainer | √√ | √√ | √√ | |
| DVC-D 17-88 Telfare Device | | | | √√ |
| DVC-D 17-89 Wallace Device | | | | √ |
| BOI Cal. .50 In-Bore Device (90, 105mm) | | | | √ |
| BOI Riley 20mm In-Bore Device | | | | √ |
| BOI M55 Laser Trainer (Coaxial Mount) | √ | √ | √ | |
| BOI 7.62mm Single Shot Device | √ | √ | √ | |

√ Initial candidate device.

√√ Initial candidate deemed acceptable.

**Table 4.**

**Additional Candidate Devices**

| Device | Initial Candidate | Initial Candidate Deemed Acceptable |
|---|---|---|
| Chrysler Conduct-of-Fire Trainer | √ | √ |
| Wiley 17-B4 Conduct-of-Fire Trainer | √ | √ |
| DVC-D 17-4 Conduct-of-Fire Tank Gunnery Trainer (Green Hornet) | √ | |
| DVC-D 17-94 Stout Device | √ | √ |
| REALTRAIN | √ | |
| Multiple Integrated Laser Equipment System (MILES) | √ | √ |
| Main Gun Simulator | √ | |
| Tank Appended Crew Evaluation Device (TACED) | √ | √ |
| Dry Fire | √ | |

test crew or crew member proficiency. These consist primarily of the various subcaliber devices used in connection with scaled gunnery Tables I-VP (FM 17-12-2). The possibility of using any or all of the devices to simulate the model marksmanship exercises was evaluated in the next four phases of the analysis.

STEP 2 - INITIAL SCREENING OF DEVICES

The feasibility of using any of the 39 devices to test crew marksmanship was determined during technical discussions with cognizant USAARMS personnel. Two subject matter experts, working independently, analyzed each device in detail. They were asked to screen out devices with hardware or design problems that would interfere with marksmanship testing (e.g., mount instability, sight and weapon parallax).

Fourteen of the 39 candidate devices were viewed as potentially useful. Nine of these were subcaliber devices fired on scaled ranges; they consisted of Brewster-mounted M55 lasers or 5.56mm or .22 caliber rifles used on various ranges, and the Telfare-mounted .50 caliber device fired on a 1/2-scale range. The nine candidates are indicated in Table 3 by double check marks. The five additional candidates that were deemed viable are indicated by check marks in the second column of Table 4. With two exceptions (MILES and TACED) these five were judged only marginally acceptable. They were carried into the next phase of analysis, however, to provide as complete a picture as possible of different approaches to the simulation of crew marksmanship exercises. The shortcomings of devices excluded from further consideration are described in Appendix A.

Virtually no documentation was readily available that described the functional characteristics or capabilities of the five futuristic devices. Therefore, formal evaluation of these devices was not attempted. (However, the Full Crew Interaction Simulator (FCIS) and the Tank Weapons Gunnery Simulation System (TWGSS), judged informally, appeared promising.)

STEP 3 - ANALYSIS OF ENGAGEMENT CONDITIONS

During Phase 1 of the project a domain of tank gunnery job objectives or tasks was defined. In keeping with an emphasis on crew marksmanship, these objectives defined all possible ways that a variety of targets could be neutralized with the 105mm main gun, the 7.62mm coaxial machinegun, and the .50 caliber machinegun weapons of the M60A1AØS tank system. Objectives were created by combining levels of all conditions associated with hypothetical engagements. The 11 conditions and levels within specific conditions are listed in Table 5,

## Table 5. Engagement Conditions and Levels of Conditions

| Crew Member | Weapon | Firing Mode | Firing Vehicle Motion | Target Motion | Target Type | Target Visibility | Day or Night | Fire Control Instrument | Ammunition | Target Range |
|---|---|---|---|---|---|---|---|---|---|---|
| 1. Gunner | Main Gun | Battlesight | Stationary | Stationary | Thin-skinned Vehicle (TSV) | Visible | Day or Night | Range-finder, Day | SABOT or HEAT | ≤1600 meters |
| 2. Tank Commander | Coax | Precision | Moving | Moving | Tank (TNK) or Light Armored Vehicle (LAV) | Visible using artificial illumination (White Light) | Night | Range-finder, Infrared (Metascope) | HEP | ≤1000 meters |
| 3. Driver | .50 Caliber | Non-precision | Moving to-a-halt | | TSV or Crew-Served Weapon | Visible using artificial illumination, Infrared | | Tank Commander's Periscope, Day Light | BEEHIVE | 500-4400 meters |
| 4. Loader | | Range Card-lay-to-direct-fire | | | Bunker or Crew-Served Weapon | Visible using artificial illumination, Flare | | Tank Commander's Periscope, Infrared | Coax 7.62 mm | 500-3200 meters |
| 5. | | | | | Troops | | | Gunner's Periscope, Day Light | .50 Caliber Machinegun | ≤4400 meters |
| 6. | | | | | Aircraft | | | Gunner's Periscope, Infrared | | 500-1600 meters |
| 7. | | | | | | | | Telescope | | ≤3200 meters |
| 8. | | | | | | | | Infinity Sight | | 500-1200 meters |
| 9. | | | | | | | | Auxiliary Fire Controls | | ≤900 meters |
| 10. | | | | | | | | | | ≤2300 meters |

as adapted from earlier reports (Kraemer, Boldovici, and Boycan, 1975; Wheaton, Fingerman, and Boycan, 1978). These conditions and their associated levels were the focus of the third stage of analysis. Each of the 14 devices was considered with respect to each condition in an attempt to identify specific facets of hypothetical engagements that could or could not be simulated.

The results of this appraisal can be characterized in two ways. First, each of the 11 engagement conditions can be considered individually, and discussed in terms of the devices that can or cannot provide for simulation. For example, the ability of each of the candidates to represent various types of (real or simulated) firing-vehicle motion can be described. These ancillary findings are presented in Appendix B (p. 48) where the focus is on those facets or conditions of an engagement that could not be simulated on the 14 candidate devices. For the purpose of pinpointing viable candidates, however, each device may be evaluated by scanning across the various engagement conditions, thus revealing the degree to which each device suffers from relatively few or many deficiencies in simulation. These results (also discussed in detail in Appendix B) are summarized below.

Based on their ability to simulate important dimensions or conditions of tank crew marksmanship exercises, three potentially viable candidates were identified: the Telfare .50 caliber device fired on a 1/2-scale range; the Brewster M16 device used in conjunction with a 1/20-scale range; and the Tank Appended Crew Evaluation Device (TACED) used on a 1/20-, 1/2-, or full-scale range. Each of these devices provides for less than full simulation, in the sense that each requires an operational M60A1AØS tank and some form of (scaled) range facility. Because of this fact some of the costs associated with livefire testing must be borne when these devices are used. But this same relatively high degree of realism makes these devices the ones of choice. They possess the greatest versatility in the conditions that can be simulated and provide most information about crew performance.

Versatility is an important consideration from a logistics point of view. It would clearly be impractical to run crews through a number of different testing devices, each of which was specifically employed to simulate one or two particular kinds of engagements. One would simply be better off in trying to keep to a minimum the number of different devices used to simulate a broad range of exercises. The three contenders were superior to their rivals in this respect. They were also superior in terms of exercising the driver, gunner, and tank commander, as well as the loader, provided this crewman was furnished with dummy rounds. Full-crew interaction of this type is vital if the simulated test is to yield valid estimates of a crew's performance relative to crew qualification standards.

## STEP 4 - ANALYSIS OF MARKSMANSHIP EXERCISES

Information generated in the preceding step was used to evaluate the capabilities and limitations of the three devices still considered to be viable candidates. The analysis focused on their potential for simulating specific gunnery exercises drawn from three different sources. The first set consisted of the 28 engagements developed during Phase 1 of the project (Wheaton, et al., 1978). These exercises constitute a model livefire test of crew marksmanship inasmuch as they are the end product of a systematic process of test development that stressed representativeness (by providing for sampling of a wide variety of engagement conditions) and generalizability (by providing for coverage of virtually all crew member behaviors). The second set included the 22 exercises comprising the Table VIII test currently used by the Army to determine tank crew gunnery qualification (FM 17-12-2). The third set was made up of the 21 exercises contained in a draft revision of the current Table VIII that emphasizes multiple engagements in tactical gunnery (Draft Change No. 2, FM 17-12-2, 1978).

Exercises from the latter two gunnery qualification tables were included in order to provide as thorough an analysis as possible. Broadening the base of test exercises was intended to make the evaluation less dependent on the specific set of engagements under consideration. This expansion was viewed as especially important to the extent that the model test might include exercises representing rather unusual or infrequently encountered engagement conditions.

The very first outcome of the analysis was the decision to limit detailed evaluation of the candidate simulators to main gun exercises only. The preceding examination of engagement conditions demonstrated clearly and convincingly that neither coaxial nor .50 caliber machinegun exercises are particularly well-suited to simulation with the three devices (or for that matter most of the others) under consideration. As a consequence, the recommended strategy for testing crew marksmanship performance on these engagements is to intersperse them with simulated main gun exercises, but to livefire them. Given that the three remaining simulators require some form of range facility, the livefiring of machinegun exercises on these same or slightly larger ranges would seem more cost-effective than either foregoing such engagements entirely, or considering special-purpose devices that might be developed for their simulation. Livefiring would be required for the 15 machinegun exercises in the model test (see Table 2), or for the 10 and nine machinegun engagements included, respectively, in the current and revised Table VIIIs.

The main gun exercises that underwent scrutiny are presented in Appendix C (p. 54). Exercises from each of the three gunnery tables are listed in Tables C-1 to C-3 and are described in terms of the 11 specific conditions of engagement. For exercises drawn from the model test (Table C-1) these descriptions are precise. They are less precise for engagements comprising the two Table VIII tests because the relevant documentation does not specify nor does the tactical testing philosophy require that each engagement be carried out in a particular way. (In this sense neither Table VIII test provides an acceptable measure of crew marksmanship since choice of firing mode, fire control instrument, etc. is ultimately left to the discretion of the crew in response to the tactical situation.) To deal with the ambiguity inherent in these cases, all of the specific exercises that a particular engagement might actually represent were listed as options in Tables C-2 and C-3. All of these alternatives were evaluated in terms of how amenable they were to simulation.

In summary, the results of the analysis are fairly clear cut. The Telfare-mounted .50 caliber device on a 1/2-scale range apparently can be used to simulate any main gun exercise associated with the three gunnery tests. This is an impressive outcome considering the range and diversity of exercises included in the model test and the two Table VIIIs. Next best is TACED. The device appears particularly well-suited to the simulation of daylight engagements and may be capable of handling a broad range of those fired at night. This nighttime use will depend ultimately on the device's ability to cope with engagements fired under infrared and low levels of illumination. In contrast, the Brewster-mounted M16 device fired on the 1/20-scale range is clearly inferior to the two preceding devices. Its primary weakness for testing crew marksmanship, is the inability of the tank commander to provide range data in support of precision engagements.

Given these findings, the Brewster M16 device was dropped from the final step in the analysis. (It should be kept in mind for the future, however, in the event that a rangefinder is developed for use on the 1/20-scale range). TACED and the Telfare device emerged as the prime candidates and were subjected to an analysis of the behaviors involved in crew gunnery.

STEP 5 - ANALYSIS OF BEHAVIORAL ELEMENTS

The final step in the analytic procedure was to obtain estimates of the quality of simulation provided by TACED and Telfare. This evaluation was the logical conclusion to a process that had begun with the identification of candidate

devices and had proceeded to evaluate them with respect to conditions of engagement and specific marksmanship exercises.

The rationale for this last analytic activity was predicated on the tank crew gunnery job-objective/behavioral-element domain specified during Phase 1 of the project. The domain consisted of a specification of the 266 ways in which an M60A1AØS crew could neutralize targets. Correspondingly, each of these exercises was described in terms of 114 specific driver, loader, gunner, and tank commander activities (see Appendix A of Boldovici, Boycan, Fingerman, & Wheaton, 1979 for detailed specification of the domain). Given this data base, therefore, the obvious question to ask is how well the candidate devices provide for simulation of each of the behavioral elements comprising a given marksmanship exercise.

The answer was provided by ratings on a three-point scale. One end of the scale signified that a given behavior could be performed in the simulator essentially in the same manner as it occurred in the actual livefire setting. At the other end lay judgments that the behavior in question was not represented in the simulator. Between these two extremes lay a gray area in which the behavior might occur in the simulator, but it would be noticeably different in some respect from that occurring in the livefire situation. Thus, a judgment was made as to whether or not the behavior was at least functionally similar to that occurring during livefire. Further discussion of the concept of functional similarity, as applied to displays and controls rather than behaviors, may be found elsewhere (Wheaton, Rose, Fingerman, Korotkin, & Holding, 1976). Suffice it to say that in the present case few behaviors were judged to fall into this middle category.

The scale was applied to the behavioral elements comprising each of the 13 main gun exercises in the model test of marksmanship. These exercises were developed to represent virtually all crew behaviors associated with main gun firing and, therefore, were representative of main gun engagements at large. Consequently, exercises from the two Table VIIIs were not examined separately. Information obtained from them would have been redundant with the more inclusive array of behavioral elements associated with the model test.

The results can be reported succinctly for each device. Given the exercises listed in Table C-1, and replacing the two BEEHIVE engagements (#103, #81) with their designated alternates (#106, #69) on grounds of current policy, TACED provides for the simulation of all but two infrared engagements (subject to further device development). In the remaining 11 exercises virtually all of the component crew

behaviors are reproduced faithfully. This outcome is not surprising since the crew is in essence dry-firing an otherwise operational tank. The only exception (for TELFARE as well) is that the loader requires two dummy rounds in order to exhibit the required behaviors in the daylight HEP engagement (#97). This evaluation of TACED applies to the firing of an initial round; firing of subsequent rounds (e.g., BOT adjustment) cannot be simulated.

Telfare can also be used to simulate the model exercises. The quality of simulation, however, is slightly degraded on four exercises. In addition to the requirement mentioned above for the daylight HEP exercise (#92), one crew behavior cannot be simulated. In firing the .50 cali[illegible] simulation of this HEP exercise, the tank commander cannot and would not apply aim off in order to achieve a target hit. The three other cases (#43, #67, #113) involve precision or range-card-lay-to-direct-fire engagements of moving targets. In these instances the gunner does not lead the target in precisely the manner he would if firing SABOT or HEAT. Consequently, three of his behaviors were judged to be only functionally equivalent: "gunner applies lead in direction of target apparent motion", "gunner lays rangeline leadline at center of target vulnerability", and "gunner makes final precise lay".

Few shortcomings were found in either device. Those that were uncovered did not appear serious. As a consequence, the content validity of a simulated marksmanship test based on the model exercises and the Telfare or TACED approach was judged to be high. All of the main gun exercises could be simulated and virtually all of the underlying behavioral elements were represented in a realistic manner. Given these outcomes the final step was to propose simulated crew marksmanship tests.

STEP 6 - SPECIFICATION OF SIMULATED TESTS

In specifying simulated tests of tank crew marksmanship that could actually be used to replace livefire testing, three issues were considered. The first concerned the specific set of exercises upon which to base the test. The second addressed the choice of device, given that both Telfare and TACED had potential. The third and final issue was the scoring system that should be used to evaluate crew proficiency. Each of these issues is discussed below in the course of elaborating the model simulated test.

Test exercises. One set of exercises on which to base the simulated test consists of the 28 engagements used in the livefire version of the model test (see Tables 1 and 2). In implementing this test one would simulate the 13 main gun engagements, using one of the devices for that purpose, and livefire the 15 machinegun exercises, engaging appropriate targets on a 1/2- or even full-scale range. All of these exercises, however, were selected within the context of livefire testing; as a consequence they represent a minimal set, chosen in light of resource and cost constraints. One could improve the overall quality of the test by expanding the number of engagements on which crew performance is evaluated. The presumed increase in validity and/or reliability could be obtained in the simulated testing environment for nominal increases in cost. Accordingly, the strategy of simply duplicating the livefire test was abandoned in favor of an alternative approach in which the greatly reduced cost of simulated testing permitted an assessment of crew performance based on a larger number of exercises.

Given the desirability of firing more engagements, alternative ways of constructing the simulated test were considered. On the one hand, new exercises could be added to those in the existing test to provide even better representation of the overall gunnery domain. Toward this end the general sampling strategies used to select the basic set of 28 exercises could again be pursued (Wheaton, et al., 1978). In theory, the content validity of the resulting test would be greater than for one based on only 28 items; however, stability of performance on any single engagement would be problematic since each engagement would only be fired once. Another alternative would be to repeat the basic set of 28 exercises some number of times. Such an approach would help stabilize estimates of crew performance on each engagement, where the systematic analytic procedures had been used to choose good representatives for each family. In theory, the reliability of the test would be improved; but the breadth of coverage would remain the same. A third option, of course, would be to combine these two approaches. A choice among these options involves the resolution of many complex and subtle issues and requires empirical study (see Chapter III and Appendix E.) Because of possible resource constraints the combined approach was not pursued. Since coverage of the domain was already considered to be adequate, the preferable option was to increase stability in estimates of performance on each engagement. This was accomplished by replication of exercises as described below.

The simulated model test is composed of 54 exercises: each of the 13 simulated main gun engagements is replicated three times; the 15 machinegun engagements are livefired once. (The decision not to replicate the machinegun exercises was fairly arbitrary and based on the opinion that typical crews have little difficulty with these engagements; given sufficient resources they also could be replicated.) In an attempt to balance out carry-over and learning effects (discussed in detail below), three random orders of the seven daylight main gun engagements were developed, a process repeated for the six nighttime main gun engagements. The livefire machinegun exercises were then interspersed among the main gun engagements: during daylight firing, every third simulated main gun engagement is followed by a livefire machinegun exercise, providing for two of the latter in each of the three replications; at night the machinegun engagements are inserted after two simulated main gun exercises, providing for three livefire machinegun engagements in each replication. The simulated tank crew marksmanship test is presented in Appendix D. Daylight engagements appear in Table D-1 while those fired at night may be found in Table D-2.

Simulators. The second issue in specifying the simulated test was the choice of simulator: Telfare or TACED. Each device has obvious strengths: Telfare allows the crew to put "steel on target," yielding target effects (i.e., knock-downs, sensings) that enable the crew to adjust subsequent fire if necessary; TACED permits the crew to go through all of its required behaviors, and provides a hard copy record of the resulting proficiency as indicated by the consequent sight picture. Both have weaknesses: Telfare, like any other weapon system, is subject to dispersion effects that may penalize the crew in spite of perfect performance; TACED is of questionable value under low light conditions, and in no case can it simulate firing of subsequent rounds (since the first "round" cannot be sensed).

It is recommended that both devices be used to simulate and measure crew marksmanship, especially because of dispersion effects. TACED can pinpoint dispersion by providing a record of the final sight picture which may be compared to the strike of the Telfare round. TACED can also verify Telfare target effects. The burden of actual simulation would fall on Telfare, while TACED would improve the accuracy of performance assessment. The main gun exercises in Tables D-1 and D-2 are consequently simulated by using the two devices conjointly.

Scoring. The model simulated test is designed to provide information about crew marksmanship performance for a variety of

purposes: qualification, training diagnosis, and motivation. Scoring procedures for these purposes are detailed below. Similar procedures for the livefire model test are detailed in Wheaton, et al. (1978).

Basically, the same scoring procedures used on the model livefire test are to be used with the simulated version. Each round fired is scored either as a hit or a miss, or as a measured deviation from some idealized (e.g., center of target) aiming point. Various aspects of engagement time are measured and recorded to reflect the speed with which crews engage targets. Also recorded is the round with which a target hit is first obtained. These data, which are collected for each main gun and machinegun exercise in the test, constitute the most basic scores in a scoring system involving three hierarchical levels. Scores at the second level are derived by applying standards of performance to the speed and accuracy data obtained for each engagement. The standards (shown in Tables 1 and 2) are used to identify engagements in which crew proficiency equals or exceeds agreed upon levels (scored as a "GO") or fails to do so (scored as a "NO GO"). The patterns of "GO" and "NO GO" exercises together with the underlying performance data can be used for diagnostic purposes to identify behaviors and engagement conditions on which a crew may need remedial training.

The highest level of the scoring system results in a decision about the crew's level of competence vis à vis the overall domain of marksmanship exercises. Two aggregate scores are developed, one representing competence on main gun engagements, the other indicating competence on machinegun engagements. The part scores are calculated and evaluated against a level of competence specified separately for each type of engagement.

The three-category scheme used to determine crew qualification on the livefire test can be adopted for use on the simulated version (Wheaton, et al., 1978). This scheme is proposed only tentatively. Final standards should ultimately be developed based on an evaluation/calibration study as described in the next major section of this report. In the scheme adopted, three standards of competence are specified for main gun and machinegun marksmanship. Crews successfully performing 92% of all the repetitions of either type of engagement would be qualified on that type. Crews performing 69% or fewer would be unqualified. Crews lying between these two bounds would be viewed as marginally qualified. The overall qualification decisions that would be reached for crews, falling into the three different zones for each of the two aspects of marksmanship, are portrayed in Table D-3. A detailed discussion of this and related scoring topics appears in Wheaton et al., (1978).

The replication of main gun engagements in the proposed simulated test may be used to make the standards for qualification even more stringent. For example, an additional criterion for qualification might be added to those described above: at least one of the three replications of every main gun exercise fired must be performed successfully (in addition to meeting the 92% requirement). Such an approach to qualification emphasizes that every objective in the domain of tank gunnery is relevant, and that a "qualified" crew must be capable of performing all objectives. But it also recognizes that there are problems in measuring crew performance because of the uncertainty introduced by dispersion, whether the weapon is a main gun or a subcaliber device (thus the requirement to successfully perform one out of three, rather than three out of three replications). Finally, even if this more stringent criterion is not adopted for determining crew qualification, the idea of looking at replications of the same engagement could be of great value in scoring for diagnosis of particular crew strengths and weaknesses. Any exercise which is failed three times by a particular crew would clearly signal an aspect of marksmanship on which additional training is required.

Summary. The effort described above has resulted in the specification of a model simulated test that, conceptually at least, can be used to evaluate the marksmanship proficiency of tank crews. Development of the test is of great potential significance because of the promise it holds as a substitute for livefire testing. At this point, however, an important caveat must be raised, and a proper note of caution sounded. The simulated marksmanship test, like its livefire forebearer, has been painstakingly developed to provide a valid assessment of crew marksmanship. But the development process has been entirely analytical, and to this point has proceeded in the absence of any hard data. Such data are absolutely essential before faith can be placed in the simulated test. The next section of the report describes why this is so and discusses the approach that should be followed.

# III. EVALUATION OF SIMULATED GUNNERY EXERCISES

Preceding sections of this report have discussed why and how one would test crew marksmanship using a set of simulated exercises. The remaining issues are why and how one would evaluate such a test. While this section will focus on the specific exercises comprising the model test, the material is generally applicable to evaluation of any simulated test.

Why is an empirical evaluation necessary? The rationales used in developing both the model tank gunnery test and its simulated version are compelling; and at first glance it might seem sufficient simply to accept these rationales and to begin employing the simulated test. However, in practice it has been amply demonstrated that even tests with compelling rationales may not, in fact, measure precisely that which they were intended to measure. The simulated test is to be used in making decisions with real-world consequences including whether or not crews are qualified, and whether or not training programs are adequate. These decisions are too important to be made without high confidence in the test; it would simply be too risky to proceed solely on an analytical basis. Thus, an empirical evaluation of the simulated test is required.

## GOALS OF THE EMPIRICAL EVALUATION STUDY

In designing the evaluation it is important to consider the specific objectives of such a study. These objectives have implications for development of one or more designs capable of: answering the specific questions of interest, making explicit the necessary controls for extraneous effects stating a priori any assumptions to be made, and being sensitive to the practical constraints of experimentation in the real Army environment.

Before considering the list of detailed objectives, it is important to review the purposes of the model livefire and simulated gunnery exercises. These exercises were selected to measure the performance of tank crews exercising combat-relevant marksmanship skills. They were assembled into a test which was designed to: a) assess whether or not the crews should be considered "qualified," and b) provide diagnostic information on areas of performance in need of remedial training. It was also assumed in the development of these tests that some of the obtained performance information might be useful in making judgments about combat effectiveness (Wheaton et al., 1978). The simulated test is to be used for these same purposes. Thus, the overall goal of the evaluation is to assess the validity of the simulated test vis à vis these purposes. Since validity is constrained by reliability, the reliability of the

simulated test must also be assessed, and since scoring for qualification is criterion-referenced, scores on the simulated test must be calibrated to scores on the livefire test. These three psychometric issues and the interplay among them are discussed more fully in Appendix E. A subordinate objective is to evaluate means of enhancing performance assessment through simulated testing. These broad goals may be translated into the following list of specific objectives for the evaluation. These represent evaluation criteria against which any simulated test must be considered.

Objective 1: Assess reliability of the livefire test. The livefire test is the empirical criterion for establishing the validity of the simulated test. Ideally, therefore, the livefire test should be a perfect measure of (true-score) marksmanship proficiency. In fact, however, it is not. Despite having a strong claim for content and construct validity (Wheaton et al., 1978; Guion & Ironson, 1978) a test based on the model set of livefire exercises is likely to involve substantial error of measurement. In particular, when assessing the accuracy of gunnery (i.e., strike of the round on target) it is known that the weapon system itself adds considerable variance. Thus, even when the gunner aims perfectly, the strike of the round includes a dispersion component that is literally random and that the crew members have no control over (Brodkin, 1958; Pfleger & Bibbero, 1969; Fingerman, 1978). Such random dispersion exists in addition to biased dispersion components over which the tank crew does have control, including, for example, errors in zeroing or boresighting, and systematic variations in performance among various lots of ammunition. Therefore, when accuracy is scored, whether sensing "target" or measuring the distance between the strike of the round and the center of the target, random and nonrandom dispersion components produce errors of measurement with regard to the crew's true proficiency. When observers are used to score hits, additional sources of error are introduced that may further reduce the reliability of performance measurement. In response to such problems perfect performance on the livefire criterion-referenced test was characterized as 95% or better. In many ways this correction for dispersion and other kinds of measurement error was gross, but some type of correction was needed.

In the evaluation study more sophisticated methods for reducing error of measurement are required, since an accurate estimate of criterion performance is particularly important. The measures of livefire performance are to be used to evaluate the validity of the simulated test, and the higher the reliability of the livefire criterion measure, the higher will be the potential validity of the simulated test. Thus, it is mandatory that as many sources of measurement error as possible

be eliminated from the livefire test. Even fairly expensive means for getting accurate measurement of livefire performance should be considered. In addition to providing reliable data for the validity assessment, estimates of the cost (in terms of measurement error) of traditional versus highly controlled measurement procedures in livefire testing will be a factor in determining the trade-off between simulated and livefire testing.

Three different kinds of reliability should be assessed for the livefire test. First, there is the reliability of the qualification decision. If crews are to be labeled "qualified" or "not qualified" by the livefire test, then two kinds of measurement error can be anticipated. In the first, crews whose true proficiency is sufficient to make them qualified may nevertheless fail to pass enough exercises to be so classified; cases in which qualified crews fail the test are termed "false negatives." In the second, crews whose true proficiency is not sufficient to qualify might nevertheless pass enough exercises to be classified as qualified; such cases are called "false positives." The probability of these two kinds of measurement error may be predicted. Wheaton et al. (1978) provide an extensive discussion assuming a binomial distribution of errors; others have used more complex assumptions (c.f. Hambleton, Swaminathan, Algina, & Coulson, 1978). At a minimum, the empirical evaluation should permit a test of these theoretical predictions of classification error. If the obtained data fit these theoretical models well, as determined by goodness of fit tests for example, then little need be done to improve the reliability of the livefire test. If good fits are not obtained, then improved control of the test environment is required, and corrections for attenuation of validity coefficients (due to unreliability in the criterion) may be required.

In the discussion above, the only matter of concern is whether or not crew performance can be classified reliably with respect to some cutoff score, presumably expressed in terms of the proportion of exercises passed. However, a second kind of reliability which must be assessed is the reliability of the proportion-correct test score itself, independent of a particular standard for qualification. Reasonably high reliability must be demonstrated for this score since the consistency of the qualification decision will be heavily dependent upon it. Similarly, a reliable proportion-correct score will be useful in providing diagnostic information for training purposes.

The third kind of reliability which must be considered is the reliability of measurement on individual engagements; that is, the reliability of crew performance on the individual exercise in terms of gunnery accuracy and speed. Once again, the reliability of these measurements underlies the reliability of the qualification decision. In addition, if such detailed measurements are reliable, they will contain considerable

diagnostic information. It should be possible from such information to provide detailed feedback to crews and to unit training personnel on specific performance deficiencies.

Objective 2: Assess reliability of the simulated test. The issues described above also apply to simulators. Sub-caliber devices are subject to dispersion effects; although such effects have not been well quantified they may be more severe than for the main gun. Alternatively, the overall reliability of the simulated test may in fact turn out to be somewhat better than that of the livefire test since certain sources of distraction will be missing.[1] In either case, interest in the reliability of the simulated test revolves around reliability as a limiting factor on validity. As above, three kinds of reliability need to be assessed: a) reliability of qualification decisions; b) reliability of the proportion-correct score; and c) reliability of measures of proficiency for individual engagements.

Objective 3: Assess the validity of the simulated test. The final primary objective of the evaluation must be an assessment of the consistency with which the simulated and livefire tests measure the performance of tank crews. Three kinds of validity are of concern. The first addresses the extent to which a qualification decision based on the simulated test is consistent with a qualification decision based on the livefire test. It may be that inconsistencies in qualification decisions based on the simulated vis à vis the livefire test may be removed by using different cutting scores on the simulated test, or even radically different scoring approaches. Therefore, more creative scoring procedures for the simulated test may be investigated for their impact on validity. Accordingly, this portion of the validity assessment may necessitate an examination of methods for test calibration. The issue is discussed in Appendix E.

The second kind of validity concerns the proportion-correct score. For training-diagnostic and motivational purposes this score has many norm-referenced applications, leading to an ordering of crews based on their performance. The consistency with which the simulated test orders crews vis à vis the livefire test must be determined.

---

[1] Sources of distraction associated with livefire may sometimes be relevant to gunnery proficiency assessment (e.g., the flash/bang effect). Nevertheless, removing them should tend to improve each crew's consistency of performance, and thus the reliability of the simulated test. See Gagné (1954) for an earlier discussion of this topic.

The final form of validity is at the individual engagement level. For individual crew-diagnostic and feedback purposes one must be able to tell crews precisely what engagements they had difficulties with. If it can be shown that difficulties on particular simulated engagements are predictive of difficulties with particular livefire engagements, then the simulated test will serve this purpose quite well. The three kinds of validity are arranged hierarchically, the validities of proportion-correct and individual engagement performance scores clearly underlying the validity of qualification decisions.

Considered jointly, the three objectives discussed above define the primary purpose of the evaluation--to determine the effectiveness of simulated exercises as substitutes for livefire in crew gunnery testing. Other less critical questions might also be addressed in the same study. The next objective is one example.

Objective 4: Explore different amounts of simulated testing. One of the chief advantages of simulated testing is reduced cost. This makes it possible to try to improve the precision of test information by using a longer test (more exercises) or by replicating a short test (giving it more than once) since, in general, the longer a test the more reliable it will be.[2] We have recommended replicating engagements twice or even three times in order to obtain more precise information for qualification and performance diagnostic purposes. In the course of the evaluation cost trade-off functions could be computed comparing improvements in reliability of the simulated test scores (and hence improvement in potential validity) to the cost of additional replications.

In order to reliaze any of these objectives, the evaluation must be carefully designed. In addition to considerations of scientific rigor, the costs of the required field research must also be taken into account. The next section develops two candidate experimental designs that have been developed to satisfy the four objectives.

---

[2] This is true assuming that item error variances are homogeneous, measurement errors are uncorrelated, and item content is homogeneous. Study of the relation between test length and reliability has generally followed two tracks, one theoretical and the other empirical. In the theoretical analyses it is assumed that the test content is homogeneous, and that the additional items measure the same underlying factor as the original items. Empirical evaluations have also generally focused on tests in which it is easy to assume content homogeneity, and assumption that may not hold in a tank gunnery test, even within more restricted item domains or families. Thus, in the present case improvement in reliability was attempted with replications of the simulated exercises as opposed to the addition of (potentially non-homogeneous) items.

## EXPERIMENTAL DESIGN

The evaluation can be based on either of two kinds of experimental designs. The first kind, termed "non-replicated," is distinguished by the fact that each of the two tests, livefire and simulated, is taken once by each crew. While this design is inexpensive, it provides only the minimum of information required about reliability and validity. The second kind, termed "replicated," involves administration of one or both of the tests several times to each crew. For the increased cost one gains considerably more information on test reliability, and is able to explore the value of increasing the length of the simulated test for (later) routine administration.

Many other designs were considered initially. Some turned out to be elaborations of the designs proposed below that provide additional or more specific information on certain questions. Others were rejected as inappropriate, even though they may have frequently appeared in the psychometric literature. The reasons for some of these rejections will be raised in the discussion of the proposed designs.

Non-replicated designs. The non-replicated design is the simpler of the two types. In an evaluation based on this design, each tank crew would fire the 28 basic engagements of the simulated test once; they would then fire the same engagements once in the livefire test environment. Since neither the simulated nor the livefire exercises are replicated, test-retest and parallel-forms methods of assessing reliability cannot be employed (see Appendix E). Reliability would instead be assessed with one of the measures of internal consistency such as the split-half reliability coefficient or the alpha coefficient. Since the exercises comprising the test may not be homogeneous (e.g., main gun and machinegun exercises), the test would be broken down into at least two components for reliability assessment. Specifically, the reliability of the main gun and machinegun portions of the test would be assessed independently for both the simulated and the livefire versions.

The order in which the engagements are fired is important because there may be large carry-over and learning effects over the course of firing 28 exercises. The specific design employed must control for these carry-over effects. One way to do so is illustrated in Table 6 for the daylight portion of the test. The steps in deriving Table 6 were as follows. The main gun daylight exercises were randomly ordered as were the machinegun daylight exercises; the daylight portion of the test was then constructed by alternately taking one main

**Table 6.**

**Ordering Engagements for Control of Carry-Over Effects**

(Daylight Exercises of Model Test)

| Exercises: | |
|---|---|
| | MG # 6 |
| | mg # 3 |
| | MG # 2 |
| | mg # 5 |
| | MG # 3 |
| | mg # 6 |
| | MG # 4 |
| | mg # 1 |
| | MG # 1 |
| | mg # 2 |
| | MG # 5 |
| | mg # 4 |
| | MG # 7 |

MG = Main Gun

mg = machinegun

gun engagement and one machinegun engagement.[3] Thus any adjacent main gun/machinegun pair is preceded by approximately the same amount of prior practice. The same procedure would be followed for the nighttime portion of the simulated table. Main gun and machinegun engagements would be randomly ordered and then drawn in alternation to construct the actual test. Precisely the same order of engagements would be used for the livefire and simulated tests.

In executing this design, each tank crew should fire the daylight and nighttime portions of the simulated test on Day 1, and the daylight and nighttime portions of the livefire test on Day 2. This is preferable to having a crew fire simulated and livefire portions on the same day, since interspersing simulated and livefire testing in this fashion would contaminate estimates of reliability for each component. One might find in this non-preferred design, for example, that the reliability of the daylight and nighttime portions of the simulated test differed. In this case one would be unable to determine whether the difference was due to daylight vs. nighttime conditions, or to the interspersing of livefire exercises. By presenting the simulated and livefire tests on two different days, one would hope to minimize learning or carry-over effects from the simulated engagements to the livefire engagements. Further, by keeping the delay interval reasonably short (i.e., one day), it would still be reasonable to assume that the true proficiency of each crew has remained the same from one test to the other. If the two tests were separated by a longer period (e.g., a month), it would be more difficult to assume true score stability, a prerequisite to the assessment of validity.[4]

The non-replicated design is subject to the problem of time-bounded inferences. Specifically, the (prior) simulated test is to be compared to the (posterior) livefire test; that is, the simulated test is being used to predict the livefire test score. Alternatively, one could present the livefire test first, followed by the simulated test, and determine whether the simulated test reproduced the measures of proficiency established in the preceding livefire test. The choice is between prediction of livefire scores and "postdiction"

---

[3] An even better procedure during the evaluation study would be to use several random orders, randomly assigned to crews. One approach would be to employ a modified graeco-latin square arrangement (Myers, 1972). This procedure would permit an examination of carry-over effects independent of specific engagement types (the ordinal-position-in-sequence-effect--see Myers, p. 282).

[4] This assumption is only reasonable in some contexts. See the discussion on crew experience that is provided later in this section.

of livefire scores. Strictly speaking, neither ordering is entirely satisfactory. The goal is to determine the substitutibility of the simulated test for the livefire test or its concurrent validity. With either ordering of the tests carry-over effects may occur. Such carry-over effects could obscure the validity of the simulated test or, at the very least, make any attempt to calibrate the simulated test extremely difficult (see Appendix E for a discussion of test calibration issues).

It is not possible to have a literally concurrent test, that is, a simulated and livefire test at exactly the same time. One approximation, however, would be what is known as a "revolving door design." Each engagement would be fired twice in succession, once simulated and once livefire, thus interweaving the two tests. One could also counterbalance, firing engagement #1 simulated, then engagement #1 livefire, then engagement #2 livefire, then engagement #2 simulated, and so on. However, as discussed above, the constant interspersing of livefire and simulated engagements would make assessment of reliability extremely tenuous. Since the reliability issue is ultimately more critical than the calibration issue, and perhaps more sensitive to the specifics of experimental design, the revolving door approach has been rejected. The Day 1 - Day 2 approach, while potentially increasing the difficulty of calibrating the simulated test, should minimize simulated-to-livefire carry-over effects and thus insure adequate assessment of reliability. One improvement in the basic design would be to split the sample crews into two groups. One group would receive the two tests in the order: simulation followed by livefire; the other group would receive the two tests in reverse order. This design is layed out in Table 7. The symbolic notations in Table 7 correspond to the variety of reliability and validity coefficients which could be obtained. For example, $r_{s1}$ refers to the reliability of the simulated test on Day 1.[5] Another example is $r_{L2}$; this refers to the reliability of the livefire test on Day 2. By comparing, for example, $r_{s1}$ to $r_{s2}$, we may assess the change in reliability as an impact of carry-over effects; $r_{s1}$ is the reliability of the test with no test-specific prior experience, while $r_{s2}$ is the reliability of the simulated test following a similar livefire test. If these two coefficients are not significantly different, the

[5] Actually there are four sets of reliability coefficients represented by $r_{s1}$. They are: the reliability of the main gun portion of the simulated test during daylight; the reliability of the main gun portion of the simulated test at night; reliability of the machinegun portion of the test during daylight; and reliability of the machinegun portion of the test at night.

**Table 7.**
***Counterbalanced Arrangement for the Non-Replicated Design***

| | Day 1 | | Day 2 | |
|---|---|---|---|---|
| Group 1: | | | | |
| $r_{S1}$ | Simulated / Daylight | V pre | Livefire / Daylight | $r_{L2}$ |
| | Simulated / Night | | Livefire / Night | |
| Group 2: | | | | |
| $r_{L1}$ | Livefire / Daylight | V post | Simulated / Daylight | $r_{S2}$ |
| | Livefire / Night | | Simulated / Night | |

data may be pooled to gain a more precise estimate of reliability of the simulated test. If they are very different, the reliability estimate on Day 1 would be more instructive since it would reflect performance in the absence of prior test-specific experience. Similar comparisons are possible for the reliability coefficients associated with the livefire test.

Two validity coefficients are also shown in Table 7. One, labeled $V_{pre}$, is for the simulated-test-before-livefire test; the other, labeled $V_{post}$, is for the simulated-test-after-livefire test. Again, if these coefficients are of similar magnitude the data may be pooled to get a more precise estimate of the validity of the simulated test. If they are very different, however, subtle distinctions in their interpretation may be appropriate. The $V_{pre}$ coefficient is a measure of the predictive validity of the simulated test, that is, how well the simulated test predicts subsequent livefire performance. The $V_{post}$ coefficient may also be interpreted as a predictive validity coefficient, that is, how well livefire performance predicts simulated performance. In cases where the two coefficients differ, the former $V_{pre}$ index would generally be preferred.

If no significant carry-over or learning effects are obtained (i.e., $r_{s1} = r_{s2}$, $r_{L1} = r_{L2}$, $V_{pre} = V_{post}$) interpretation is straightforward. When the coefficients reveal significant carry-over effects, the fine structure of the data may be revealing. For example, one may be able to determine that some engagements reveal strong carry-over effects while others do not. Further separate analyses might focus on engagements which do and do not exhibit carry-over effects.

If validity of the qualification decisions is deemed high, then computation of a calibration formula would follow via regression analysis. Rather than addressing whether or not the two tests ordered crews consistently, this analysis would determine whether or not the same qualification decisions were dictated by the two scores (livefire and simulated). The intercept parameter from the regression analysis would, in fact, be the calibration value, and its significance could be tested using standard methods. If validity at the qualification-level were low, then calibration would be more complex, depending, for example, on whether validity using finer measures of performance was low or high.

The richness of even this simple design is illustrated by the large number of reliability and validity coefficients which would have to be considered. Fortunately, not all of these coefficients are independent, and a synthesis of the results should be possible. In many cases simple inspection

of the pattern of results will enable one to make some general statements regarding the reliability and validity of the simulated marksmanship test. In other instances it may prove necessary to employ sophisticated analytical methods to achieve a synthesis.

If the estimated validity of the simulated test is high, no further comment is required. If the validity is low, two possible problems should be examined. First, one should question whether the reliabilities obtained for the simulated and livefire tests are sufficiently high. Were reliability on the livefire portion of the test low, then concerns regarding errors of measurement would be well founded, and if all countermeasures were taken (see implementation guidelines), little recourse would be available.[6] The same would be true were reliability of the simulated test found to be low. The second possible problem, suggested by a low estimate of validity but high reliability on both tests, is that during development of the simulated test some critical components of the task having an impact on performance may have been overlooked. These components might be intrinsic to the task, such as particular behavioral elements required in the livefire but not in the simulated environment, or others demanded by the simulation but not found in the livefire version. Alternatively, the critical components might be extrinsic features of the livefire environment that are missing from the simulation (or vice versa). An example is the noise and recoil (flash/bank) associated with service firing of the main gun. If the simulated and livefire tests are suspected of differing in terms of such critical features, then these factors should be identified and a new version of the simulated test should be constructed where such distinctions are removed. This step would improve the test's (content) validity. In the present case the problem of high reliability but low validity is considered highly unlikely because of the effort made to insure the content validity of the simulated test (Guion & Ironson, 1978).

Replicated designs. In this class of designs at least one of the tests is administered more than once, several replications being likely. The design is a good one because it can be used to determine whether the theoretical gains associated with increasing test length are actually realized. Because of the cost of livefire testing it is unlikely that there would be interest in routine administration of a replicated livefire marksmanship test. However, for the evaluation study, such replications would be valuable.

The chief problems with replicated designs are the same as those found in non-replicated designs: carry-over effects and time-bound inference considerations. Carry-over effects

[6] This is true assuming that item error variances are homogeneous, measurement errors are uncorrelated, and item content is homogeneous.

are more complicated here since, in addition to engagement-to-engagement and simulated-to-livefire effects, there now are replication-to-replication carry-over effects. Individual engagement carry-over effects (e.g., across main gun and machinegun exercises) can be dealt with in the same fashion as in the non-replicated designs (see Table 6). The replication-to-replication carry-over effect is more complicated, since it depends on the precise shape of the learning curve. In general, the best way to resolve this problem will be to conduct a pilot study. One group of subjects would be exposed repeatedly to the simulated test (to study learning effects in that context), and a second group would receive the livefire test several times. Design of the pilot study is straightforward. Engagement carry-over effects would be controlled for as in Table 6. Replications should not immediately follow one another; a moderate delay between replications is required, on the order of half a day to one day. Troops should have the same background, prior experience, and training as the troops to be employed in the main study. Most importantly, their experience immediately prior to the test should be the same as that of crews to be used in the main test (see the section below on implementation guidelines). Extensive prior training, e.g. Tables I through VII, is likely to reduce learning effects across replications of the tests.

Precise design of the main evaluation study will depend on results of the pilot study. These most likely will indicate that repeated administrations of the simulated test result in improved performance but, hopefully, not to the extent that severe ceiling effects are encountered. If the pilot study covers an appropriate range of replications (e.g. up to seven or ten), the data might be used to select that number of replications which improves the reliability of the simulated test without leading to ceiling effects. The same number of repetitions would then be employed in the evaluation study. Performance scores would be collapsed across replications to characterize crew proficiency.

Results of the livefire pilot study could be somewhat different. Because of the flash/bang effect, early livefire performance is likely to suffer; as crews adapt to this effect, performance will improve rapidly. This will result in a steep learning curve initially, with substantial improvement over the first two or three replications (as crews adapt), followed by a slower continuing growth due to more typical learning effects. The number of livefire replications to be used in the evaluation study could be chosen in the same way as the number of simulated replications (i.e. that number which leads to improved reliability but which precedes the occurrence of ceiling effects). However, a two-stage

rise in the livefire learning curve would complicate matters. A rapid change in scores over the first two or three replications would not, in fact, reflect true proficiency. A more desirable measure, therefore, might be a composite across replications four to six where performance would still be improving, but at a more typical rate. This measure would presumably be more stable and hence more reliable.

The possibility of carry-over effects from the simulated test to the livefire test and vice versa also exists in the replicated design. The effect may again be dealt with by counterbalancing treatment order across two groups of subjects. Repeated administrations of the simulated test, one a day, would be followed by administrations of the livefire test for one group, and vice versa for the other. The time-bound inference problem is handled by this same counterbalancing (see Table 7).

Implementing a replicated design is considerably more complex than implementing a non-replicated design. The cost of the study is also increased, not just merely because of the replications, but because of the need for pilot studies. Both designs provide information on the basic issue, the validity of a simulated test as a substitute for livefire testing. The replicated design may provide a better estimate, since it attempts to maximize the underlying reliabilities, and therefore the potential validity. The key to using this design lies in the pilot study. By running this study first, one may estimate the contribution of replications to increased reliability. If replication does not improve reliability (an improbable outcome) then validity is not likely to improve either, and the non-replicated evaluation design is appropriate. If replication does improve reliability, then the replicated approach is justified despite its greater cost. The recommendation, therefore, is that the pilot study be conducted before a final evaluation design is selected.

Number of subjects. In general, the power or sensitivity of an experiment depends to a great extent on the number of subjects tested. In the present case procedures exist for determining an appropriate number of subjects. In designing an evaluation study, interest lies in determining reasonably accurate estimates of the true (population) values of the reliability and validity coefficients, so that the risk involved in substituting a simulated for a livefire test may be determined. Thus, the criterion for a sufficient number of subjects is the width of the confidence interval for each reliability or validity statistic computed.

A confidence interval is "an estimated range of values with a given high probability of covering [including] the true population value (Hays, 1963, p. 288)." Thus, if one computes a validity coefficient from empirical data of, for example, .70, one would be interested in how close this estimate comes to the true value of the validity coefficient. A confidence interval may be computed, based on this value and the sample size, that will include the true population value at least some given percentage of the time. For example, if the validity coefficient (.70) is a Pearson product-moment correlation and is based upon a sample of 130 crews, the probability is .95 that the true validity coefficient is at least .60, and not more than .778. If the sample size is reduced to 50 (with the same obtained correlation of .70), the .95 confidence interval ranges from .534 to .818; if the sample size is further reduced to 25, the confidence interval expands to include values from .412 to .858. Since a validity of .412 is not useful in a gunnery testing context (e.g. only 16% of the variation in livefire crew performance would be predicted from the simulated test), these differences in the confidence intervals for the same validity estimate are quite critical.

The procedure for determining the required sample size is to specify a desired or expected validity or reliability coefficient value, a minimum value which would still be useful, as well as the maximum likely value, and to compute the sample size which would be required for a confidence interval which contains the expected value and is anchored by the other two values.[7] This procedure requires some guesswork in advance of the experiment, but there is no alternative. Furthermore, if subjects are grouped (e.g., for counterbalancing) and coefficients are computed within groups, the sample size calculated via this method is for the group, not the entire pool of subjects.

Reduced level of effort. Several ways of reducing the level of effort required in the evaluation may be possible, while several others should be avoided. A reduction in the level of effort is available if machinegun exercises are not evaluated. In the simulated model test machinegun exercises

[7] Formulae for computing confidence intervals depend on the specific coefficient being studied. The evaluation study is likely to require other coefficients in addition to the Pearson correlation (e.g. tetrachoric correlation, Kappa coefficient, etc.). Formulae for their confidence intervals are available in standard statistical reference works.

are livefired rather than simulated. Therefore, such exercises could be eliminated from the evaluation study. The amount of time required per crew in either of the evaluation designs would be significantly reduced by this tactic. The trade-off is that information on the reliability of machinegun exercises is lost.

Another method of reducing the level of effort would be to evaluate the concept of simulated testing on a sample of main gun exercises rather than on the entire set. For example, simulation of perhaps four main gun daylight engagements (out of seven) and four main gun night exercises (out of six) might be evaluated. If the exercises were appropriately sampled, by choosing them to represent the range of behaviors and conditions incorporated in the test, then the results of the evaluation still might generalize to the entire set of exercises.

Certain other methods for reducing the level of effort should not be used. For example, an inconvenient feature of the proposed designs may be the recommended calendar time for the study. It would not be a good idea to attempt to greatly compress the time over which each crew fires the full set of study exercises; the intervals recommended above should be maintained in order to forestall, as much as possible, the complications of carry-over effects. In an attempt to reduce the duration of the experiment arbitrary modifications in the sequencing of test components should not be undertaken. Given that the particular sequences have been based on empirical data from pilot studies, or on well-founded assumptions, changes in these sequences may lead to uninterpretable results. Finally, the level of effort should not be manipulated by arbitrarily changing the number of subjects required in the study. If the number of tank crews is reduced, the cost of the study will still be substantial, with no clear-cut empirical outcomes regarding the study objectives.

## IMPLEMENTATION GUIDELINES

Several aspects of the conduct of the evaluation are critical to its success, independent of specific experimental designs or methods of simulation. These aspects include: the crews selected to serve as subjects, measurement procedures, and control of the study.

The subjects used in the evaluation will be tank crews. As it will be necessary to generalize from results of the evaluation to tank crew testing in general, it will be important that selected crews be representative of the diversity of tank crews in the Army. Thus, they should be heterogeneous with regard to background, training, and experience. Results of the evaluation would be of little value were a single homogeneous set of crews used, for example, either AOB students

or master gunners. The extent to which the subject crews are broadly representative of crews in the Army will determine the utility of evaluation findings.

The long-term experience of the crews must be taken into consideration, together with their experience immediately prior to the test. For example, it is assumed that crews will be tested on Table VIII after a series of training/testing exercises on Tables I through VII. The extensive practice afforded by these seven tables is desirable, since it should help minimize carry-over and learning effects during the actual evaluation. (Most of the short-term learning will have taken place prior to the start of the test.) While this procedure may suppress some performance variance among crews, it is representative of the typical qualification situation. Thus, it is recommended that the evaluation be conducted in the context of the normal training and testing program for tank crews. In particular, the study should not be conducted using crews who have not recently practiced their gunnery skills.

A second set of guidelines has to do with measurement. First, there is the problem during livefire testing of measuring gunnery accuracy. The use of observers alone to determine target hits or misses is not sufficient. Data collected by observers may contain too little detail to be satisfactory for the evaluation. Further, the likelihood of observer errors is high. Obscuration, noise, and small targets may degrade their ability to score hits and misses accurately; since it is difficult to maintain their motivation, even minor distractions may become a problem. A better procedure would be to physically measure the strike of the round relative to the center of the target. This measurement could be accomplished in a number of ways, including physically measuring and "pasting up" targets, or using video cameras equipped with telephoto lenses stationed in an overwatch position. Tapes obtained by this means could then be displayed on a monitor for scoring purposes. The use of these kinds of procedures should not necessarily preclude the use of observers. Among other things, the use of observers in conjunction with one of the other recommended procedures would permit an empirical assessment of the error of measurement introduced by observers, a study never before undertaken.

Another problem in determining the accuracy of gunnery is the deviation between a gunner's aiming performance and the actual strike of the round. As mentioned above, dispersion effects exist that lead to differences between the gunner's sight picture on the target and the actual strike of the round. Since it is the gunner's performance which is of primary interest, it will be useful in some cases to factor

out these dispersion effects. One way of doing so is to record the actual sight picture at the time of firing using a video gun camera system (e.g. TACED) and to compare these data to strike-of-the-round information. Previous comparisons of this type have proved quite valuable (Fingerman, 1978). Use of the video gun camera system would permit explicit scoring of the sight picture, a key indicator of the crew's gunnery performance, independent of any weapon system effects.

Measurement of gunnery accuracy is also of concern in the simulated test. TACED should be adequate since the video image explicitly represents the firing crew member's sight picture. When subcaliber devices are used in the simulated test, there are several alternatives for measurement of gunnery accuracy. The simplest is the use of knock-down scaled targets. An observer should certainly be consistent in determining whether a target has been knocked down (compared to whether or not a main gun round has struck a target), although there is still potential for error of measurement. Short rounds, for example, may knock down scaled targets, and such misses may be indistinguishable from actual hits.

Another aspect of measurement which must be considered is the speed of engagement. A minimal requirement would be to equip observers with stop watches, and link the observers into the tank communication system. They would time the engagement by starting their watches upon hearing the actual fire command, and stopping their watches upon firing of the round. Since two-round engagements are included in the test, a stop watch with a split sweep hand would be employed. Such watches can be used to time two consecutive intervals. The observer would start timing upon hearing the initial fire command (or upon presentation of the target), stop it when the first round is fired, and stop it again when the second round is fired. The two hands would then indicate time-to-fire the first round, time-to-fire the second round, and the time between the first and second rounds. Time data might also be measured from the sound track of gun camera tapes. This soundtrack can be used to record communications on the tank intercom system, and rounds which are fired can clearly be heard and seen on the video tape. Since the scoring can be done after testing, in a quiet and controlled environment (e.g. an office), this method may be more reliable. If sophisticated equipment is available, clock times can actually be superimposed on the video image, and read directly from the tape as recorded events are played back. Such a system has been used recently by the Armor Engineer Board at Fort Knox (Fingerman, 1978).

The last set of implementation guidelines has to do with the control and standardization of procedures during the evaluation. The first issue, and perhaps the most critical, is control over the manner in which exercises are fired. Each exercise must be fired by each crew exactly as specified. That is, if a tank-commander/battlesight engagement is called for at a particular time, control personnel must insure that the tank commander does in fact fire a battlesight engagement. Crews must not be free to select the method of engagement or the crew member who fires. This is particularly critical in the evaluation context, since individual simulated exercises will be compared to individual livefire exercises. If the same crew member has not fired the same exercise using the same method of engagement, obtained reliability and validity coefficients will be uninterpretable (since they will be based on unmatched exercises).

Similarly, other control procedures must be standardized. Evaluators and control personnel should receive formal training in methods of data recording and data control, and in methods of controlling the exercises under field conditions. Instructions provided to tank crews should be standardized and read by test administrators. Forms should be prepared for recording data. Schedules should be established for running the study, and every attempt should be made to insure that a crew fires on schedule. In cases where a crew is tested more than once (day and night, or on two consecutive days) efforts must be made to insure that all crews actually appear for each test session. If a crew does not complete all of the test exercises as scheduled, the remaining data for that crew will have to be discarded.

The final aspect of control which must be considered deals with ambient environmental conditions. Of special concern are poor weather conditions such as rain, snow, sleet, fog, or high winds. Changes in the physical characteristics of the firing ranges used, or failures in tank weapon systems are also important. At a minimum such environmental conditions must be the same for all firing done by any single crew (with the exception of illumination differences for day vs. night). For example, were simulated exercises fired under clear weather while livefire exercises were fired under rainy conditions, differences in performance under the two conditions could not be clearly attributed to the differences between simulated and livefire engagements. The simulated vs. livefire comparison would be confounded with changing weather conditions, and the data would be useless for calibrating the two kinds of tests. Ideally it would be desirable to have all crews fire under exactly the same environmental conditions, especially when groups of crews receive different treatments (as in the nonreplicated design, Table 7). If different

groups of crews were tested under different environmental conditions, the group treatments would be confounded with environmental conditions, and performance comparisons would become ambiguous. The point is that as much control as is practicable must be exerted to insure that crews fire under comparable conditions.

There are several ways to control for such environmental problems. Optimally all crews should be tested during a single period when a weather forecast indicates equivalent conditions across the test period. If all crews cannot be tested within a two-day period, for example, then different groups of crews should be tested in a counterbalanced order. Thus, assuming that a full test required two days (e.g. Table 7), some portion of the Group 1 crews and some portion of the Group 2 crews would fire during the first two-day cycle (thus controlling for weather conditions for these subgroups), additional portions would fire in a second two-day cycle, and so on until all testing had been completed. With this kind of counterbalancing, even if weather conditions varied from the first two-day cycle to the second two-day cycle, the impact would be equated across experimental conditions.

Finally, all tank systems should be rigorously maintained during the course of the study. Boresighting and zeroing should be performed immediately prior to the start of the study, and zero confirmation should be repeated each time a crew begins a new set of test exercises.

Some might argue that these guidelines are extremely restrictive, and inappropriate to real world tank crew testing. While the desirability of such control procedures for routine testing can be debated (but see, for example, Wheaton et al., 1978), there can be no such debate for the evaluation. This study is a one-time occurrence. Strict measurement and test administration procedures coupled with a representative sample of tank crews are absolutely essential if results of the study are to resolve whether simulated exercises can be substituted for livefire testing of crew marksmanship.

## REFERENCES


Boldovici, J. A., Boycan, G. G., Fingerman, P. W., & Wheaton, G. R. Tank gunnery data handbook. Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences, 1979.

Brennan, R. L., & Kane, M. T. An index of dependability for mastery tests. Journal of Educational Measurement, 1977, 14, 277-289.

Brodkin, H. Fire control studies: Tank gunnery accuracy evaluation. Report R-1380A. Philadelphia, PA: Fire Control Instrument Group, Frankford Arsenal, 1958.

Cohen, J. Statistical power analysis for the behavioral sciences. New York: Academic Press, 1969.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements. New York: Wiley, 1972.

Fingerman, P. W. A preliminary investigation of weapon-system dispersion and crew marksmanship. Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences, 1978.

Gagné, R. M. Training devices and simulators: Some research issues. American Psychologist, 1954, 9, 95-107.

Guion, R. M. Principles of work sample testing: I. A non-empirical taxonomy of test uses. Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences, 1978a.

Guion, R. M. Principles of work sample testing: II. Evaluation of personnel testing programs. Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences, 1978b.

Guion, R. M. Principles of work sample testing: III. Construction and evaluation of work sample tests. U. S. Army Research Institute for the Behavioral and Social Sciences, 1978c.

Guion, R. M., & Ironson, G. H. Principles of work sample testing: IV. Generalizability. Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences, 1978.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Harris, C. W. An interpretation of Livingston's reliability coefficient of criterion-referenced tests. Journal of Educational Measurement, 1972, 9, 27-29.

Hays, W.L. Statistics. New York: Holt, Rinehart, & Winston, 1963.

Kraemer, R. E., Boldovici, J. A., & Boycan, G. G. Job objectives for M60A1AOS tank gunnery compared to proposed training. Vol. 1: Development and results. Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences, 1975.

Livingston, S. A. Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 1972, 9, 13-26. (a)

Livingston, S. A. A reply to Harris' "An interpretation of Livingston's reliability coefficient for criterion-referenced tests." Journal of Educational Measurement, 1972, 9, 31. (b)

Livingston, S. A. Reply to Shavelson, Block, and Ravitch's "Criterion-referenced testing: Comments on reliability." Journal of Educational Measurement, 1972, 9, 139-140. (c)

Myers, J. L. Fundamentals of experimental design (2nd Ed). Boston: Allyn and Bacon, 1972.

Pfleger, K. R., & Bibbero, R. J. The evaluation of combat vehicle fire control/gunnery systems. Report R-1937. Philadelphia, PA: Fire Control Development and Engineering Laboratories, Frankford Arsenal, 1969.

Rozeboom, W. W. Foundations of the theory of prediction. Homewood, IL: Dorsey Press, 1966.

U. S. Army Armor School (USAARMS). Tank gunnery devices. Draft Training Circular (TC 17-12-7), Fort Knox, KY, 1976.

U. S. Army Armor School (USAARMS). Tank Gunnery. Field Manual (FM 17-12), Washington, DC: Headquarters, Department of the Army, 1977.

U. S. Army Armor School (USAARMS). Tank gunnery for M60, M60A1, M60A1(AOS), and M48A5 tanks. Field Manual (FM 17-12-2), Washington, DC: Headquarters, Department of the Army, 1977.

U. S. Army Armor School (USAARMS). Tank gunnery for M60, M60A1, M60A1(AOS), and M48A5 tanks. Field Manual (FM 17-12-2), Draft Change No. 2, Fort Knox, KY, 1978.

U. S. Army Headquarters. Index and description of army training devices. Pamphlet 310-12, Washington, DC, 1972.

U. S. Army Training and Doctrine Command (TRADOC). Catalog of TASO training devices. TRADOC Pamphlet 71-9, Fort Monroe, VA, 1976.

Wheaton, G. R., Fingerman, P. W., & Boycan, G. G. Development of a model tank gunnery test. Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences, 1978.

Wheaton, G. R., Rose, A. M., Fingerman, P. W., Korotkin, A. L., & Holding, D. H. Evaluation of the effectiveness of training devices: Literature review and preliminary model. Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences, 1976.

# APPENDIX A

## DEVICES ELIMINATED DURING INITIAL FEASIBILITY SCREENING

As indicated below, seven of the 12 weapon/mount options listed in Table 3 were rejected outright, resulting in the elimination of 13 devices defined by a particular weapon/mount/range combination. The seven weapon/mount options that were judged unacceptable included the following:

M219 Coaxial Mount/7.62mm Single Shot Device

This weapon is incapable of delivering tight shot groups on any but the closest of targets. Parallax problems are pronounced and the single-shot mechanism is not in favor. The device has been replaced by a similar weapon in a Brewster mount.

BOI DVC-D 17-53 .22 Cal. In-Bore (90, 105mm) Device

This device yields a tighter shot group and is more economical than the 7.62mm coaxial machinegun fired in the single-shot mode. Although designed for use on 1/60- and 1/35-scale ranges, it is readily used only at tank-to-target distances of 58 to 62m because of parallax problems in the tank's fire control system. It has been essentially superseded by the Brewster-mounted subcaliber devices.

BOI M55 Laser Trainer (Coaxial Mount)

Parallax problems in the fire control system are pronounced and affect accuracy. As a consequence targets must be placed at the range for which the M55 is zeroed. For all intents and purposes this device has been rendered obsolete by an M55 in a Brewster mount.

DVC-D 17-85 Minitank Bracket/Rifle (.22 Cal.) Device

This device is designed for and is very accurate on 1/60- and 1/35-scale ranges. Parallax problems have been reduced by mounting the device over the gunner's sights. Ballistic characteristics of main gun rounds are approximated by a cam that allows for superelevation. But as a consequence, "...the range scale on the range drum must be changed for each target to correspond to the range of that target (TC 17-12-7, p. 35)." This restriction is severe in a testing situation where variations in target range (not to mention motion) will be pronounced.

DVC-D 17-89 Wallace Device

This ingenious mount allows the M85 cupola-mounted machinegun to simulate firing of the main gun. When equipped with the necessary single-shot device, however, the M85 cannot be operated in the automatic mode. This device has essentially been replaced by the Telfare Device.

BOI Cal. 50 In-Bore (90, 105mm) Device

While presumably accurate to a range in excess of 1000m, the in-bore mount is subject to instability. The device has been replaced by the Telfare Device.

BOI Riley 20mm In-Bore Device

This device is highly similar to the .50 cal. in-bore device described above. A special cam is used with the tank's computer to permit simulation of main gun round ballistics. Frequently, however, the special cam damages the computer. The device has been superseded by the Telfare Device.

Of the remaining 12 devices in Table 3, three more were eliminated after further analysis. Although the Brewster-mounted 7.62mm machinegun was viewed as an improvement over the same weapon in a coaxial mount, it was eliminated from further consideration on two counts. The M219 machinegun was judged not to fire reliably (a problem not associated with the M240 on the XM1 tank) and to require the balky single-shot instrumentation. This judgment precluded use of this weapon/mount on 1/20- and 1/2-scale ranges. Finally, the judgment was made that a Brewster-mounted M16 rifle firing 5.56mm ammunition would be overextended on a 1/2-scale range. It was, therefore, eliminated, bringing the total number of viable scaled-range approaches to nine.

Considering the devices listed in Table 4, four were quickly eliminated from further evaluation. The Green Hornet was viewed as a classroom training aid, useful in conveying principles of burst-on-target adjustment of fire techniques but not suited to evaluating crew marksmanship. REALTRAIN, a superb device for training and evaluating tactics, was eliminated because it did not represent a precision gunnery engagement system; target effects were lacking, creating a variety of problems for crews and evaluators alike. The Main Gun Simulator, however enticing the label, was simply a pyrotechnical device. Finally, dry fire scored by an observer was deemed imprecise, and entirely too subjective, lacking in any record of performance.

The remaining five devices in Table 4 were deemed provisionally acceptable for use in evaluating crew marksmanship or in testing specific facets thereof. In all but two of the cases the devices were viewed as marginally acceptable since they did not allow for full or even partial crew interaction. The two exceptions were MILES and TACED.

# APPENDIX B

## ENGAGEMENT CONDITIONS NOT SIMULATED ON SPECIFIC DEVICES

The findings relating specific devices and engagement conditions are shown in Table B-1. The entries in each column reflect the status of the 14 candidate devices vis a vis the specific condition in question. Labels reflect the specific facet that cannot be simulated for testing purposes; "OK" denotes an acceptable device vis á vis a given condition; "None" in the target type and ammunition columns means that no specific facet (e.g., light-armored vehicle, crew-served weapon, SABOT, HEP) is simulated. The rationales underlying these judged limitations are discussed below briefly.

None of the devices provides adequately for involvement of all four crew members in their normal, highly interactive mode of operation. Noninvolvement of the driver is particularly serious and results from the simple fact that many of the devices cannot simulate tank motion. Similarly, the loader is not exercised adequately on any of the devices as now configured. This latter deficiency can, however, be overcome by having the loader load appropriate dummy rounds on those devices that are installed in the M60A1AØS tank. All of the devices make provision for inclusion of the gunner, and most also permit evaluation of the tank commander as an active crew member.

The 14 devices under consideration are primarily main gun simulators. In no instance is the cupola-mounted .50 caliber machinegun simulated, and in only a few cases is the coaxial machinegun simulated (i.e., on the 17-B4 and M55 laser devices). An alternative, therefore, is to livefire these weapons on 1/2- or full-scale ranges while simulating the main gun on 1/20- to 1/2-scale ranges. Using this approach three devices (i.e., #7, #8, and #9) can be used to simulate all of the tank's weapon systems.

The major restriction on firing mode is the fact that precision engagements cannot be fired on the smaller scaled ranges (i.e., 1/60- to 1/20-scale). This limitation arises because the tank commander's coincidence rangefinder, from which range data vital to precision engagements are obtained, cannot be used on ranges of less than 1/2-scale.

As previously noted, many of the devices do not permit simulation of engagements in which the firing vehicle is shooting on the move or after moving to a halt. To accomodate such conditions, as one would almost certainly want to during the evaluation of M60A1AØS crews, 1/20- or larger-scale range facilities are required. On the other hand, various conditions of target motion are portrayed by most of the devices. Only the Stout and Wiley 17-B4 devices fail

## Table B-1. Engagement Conditions Not Simulated on Specific Devices

| Devices | Crew Member | Weapon | Firing Mode | Firing Vehicle Motion | Target Motion | Target Type | Target Visibility | Day or Night | Fire Control Instrument | Ammunition | Target Range |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. Brewster, .22 Caliber adapter 1/60 scale | Driver Loader | Coax .50 Caliber | Precision Nonprecision (Machineguns) | Moving Moving-to-a Halt | OK | OK | OK | OK | Rangefinder TC Periscope Day TC Periscope IR Metascope Infinity Sight Telescope | None | OK |
| 2. Brewster, .22 Caliber adapter 1/35 scale | " | " | " | " | " | " | " | " | " | " | " |
| 3. Brewster, M16 1/60 scale | " | " | " | " | " | " | " | " | " | " | " |
| 4. Brewster, M16 1/35 scale | " | " | " | " | " | " | " | " | " | " | " |
| 5. Brewster, M55 1/60 scale | " | .50 Caliber | Precision Nonprecision (M85) | " | " | " | " | " | " | " | " |
| 6. Brewster, M55 1/35 scale | " | .50 Caliber | " | " | " | " | " | " | " | " | " |
| 7. Brewster, M16 1/20 scale | Loader | OK | Precision | OK | " | " | " | " | Rangefinder Metascope Telescope | " | " |
| 8. Brewster, M55 1/20 scale | " | " | " | " | " | " | " | " | " | " | 2000+ m |
| 9. Telfare Device | " | " | OK | " | " | " | " | " | OK | SABOT HEP BEEHIVE | OK |
| 10. Chrysler COFT | Tank Commander Driver Loader | Coax .50 Caliber | Nonprecision (Machineguns) Range Card-lay-to-direct-fire | Moving-to-a Halt | " | None | White light Flare Infrared | Night | Rangefinder TC Periscope Day/IR Metascope Infinity Sight Telescope Auxiliary Controls Gunner Periscope IR | BEEHIVE | All but 1000m 1500m 2500m |
| 11. Wiley 17B4 COFT | " | .50 Caliber | Precision Nonprecision (M85) Range Card-lay-to-direct-fire | Moving Moving-to-a Halt | Moving | OK | " | " | Rangefinder TC Periscope Day/IR Metascope Gunner Periscope IR Auxiliary Controls | None | OK |
| 12. Stout Device | Driver Loader | Coax .50 Caliber | Nonprecision (Machineguns) | " | " | " | OK | OK | TC Periscope Day/IR Metascope Infinity Sight Telescope | " | " |
| 13. MILES | Loader | Coax .50 Caliber | Precision Nonprecision (Machineguns) | OK | OK | " | " | " | OK | " | " |
| 14. TACED | " | .50 Caliber | Nonprecision (M85) | " | " | " | Infrared Flare | " | TC Periscope Day/IR Metascope Infinity Sight Gunner Periscope IR | " | " |

to provide for moving as well as stationary targets (although at least one prototype of the latter device has been developed in which moving targets are presented by means of a motion picture system).

Most of the devices are quite flexible with respect to the target array that can be provided. Most target types are readily simulated with the possible exception of in-flight or hovering aircraft. For the most part, as indicated in Table B-1, targets can be engaged either during daylight conditions or at night using different kinds of artificial illumination. Here again most of the devices provide for several options including flares, whitelight (xenon searchlight) and infrared illumination.

Major constraints surround the fire control instruments that can be used in the various simulations. For example, in many cases use of the rangefinder is restricted for reasons cited earlier (although it can be used to engage targets from the tank commander's position, as opposed to ranging on them). When use of the tank commander's daylight or infrared periscope is curtailed it is usually because .50 caliber engagements cannot be conducted. In practice the metascope is simply no longer used under any engagement conditions. The gunner's telescope was deemed unusable on many of the shorter-scale ranges, even though it can in actuality be used in engagements conducted at the zeroing range. However, the inflexibility imposed by such a requirement led to the downgrading of several of the simulations.

Finally, none of the devices permitted simulation of the ballistic characteristics of all of the different kinds of ammunition, since in most cases the computer must be turned off. Best representation occurs in the Chrysler COFT where the flight times and trajectories of all main gun rounds except BEEHIVE can be simulated. On the other hand, simulation of tank-to-target ranges appeared adequate on most devices. Potentially important constraints were noted only on the Chrysler COFT (three ranges for stationary or crossing targets) and the Brewster-mounted M55 device fired on the 1/20-scale range (maximum simulated range of 2000m after which the capability of the laser is exceeded).

Returning to Table B-1, these various limitations were used to evaluate each device. Information may be aggregated across columns, and the fewer limitations a given device has, the more desirable it is as a simulator for testing purposes. On this basis, the .22 caliber, M16, and M55 devices fired on 1/60- and 1/35-scale ranges have a number of drawbacks that preclude further consideration of them. By virtue of

the fact that all six are associated with the smaller scaled ranges (i.e., 1/60- and 1/35-scale) they cannot be used to simulate those engagements in which the firing vehicle is moving; nor can they, as a consequence, provide for adequate testing of the driver's contribution to crew performance. Other major limitations related to the size of the ranges include: an inability to fire precision engagements, restrictions on the fire control instruments that can be employed, and an inability to field real or simulated machinegun exercises. The conclusion, therefore, is that none of these six devices is sufficiently versatile to be used in a comprehensive test of crew marksmanship. In particular, none can simulate the model set of test exercises that was designed to maximize variety in engagement conditions.

Brewster-mounted M16 and M55 devices fired on the 1/20-scale range are somewhat superior to the preceding six devices. The driver is fully involved because firing on the move or after moving to a halt is possible. Machinegun engagements can be fired on 1/2-scale or full-scale ranges. The major drawback is the inability of the tank commander to range with his rangefinder, thereby precluding the firing of precision engagements. After careful consideration of both 1/20-scale devices, the decision was made to retain the M16 version and to drop the M55 laser device. The latter was eliminated because of a restriction on the maximum range (2000m simulated or 330 feet actual), and because of a requirement that moving targets be engaged within a 15° arc of the tank-to-target perpendicular. The latter problem was viewed as particularly serious when attempting to simulate moving tank and/or moving target exercises, of which there are several in the model test.

The Telfare-mounted, M2 .50 caliber weapon, fired on a 1/2-scale range, was retained for further scrutiny. Its capacity for simulating most of the major engagement conditions was deemed superior to that of the other 13 candidates.

The Chrysler and Wiley Conduct-of-Fire Trainers and the Stout device were eliminated because of numerous shortcomings (see Table B-1). They simply are not sufficiently versatile (nor were they intended to be) to provide simulations of a wide variety of exercises. In those instances where they otherwise do appear to have some potential, the full crew interaction , so typical of and necessary in gunnery, is clearly absent.

MILES seems attractive from a testing point of view, provided that arrangements can be made to livefire the machinegun exercises. The attractiveness of this approach is offset by what is assumed to be the much greater cost of the system, unknown system reliability, and the fact that it

is still undergoing development and military potential testing. As experience is gained with this simulator it may indeed rival or surpass the other viable candidates. Until that time, however, the use of MILES as a marksmanship testing device should be held in abeyance in deference to other cheaper and more readily available devices.

The final candidate, TACED, deserves further consideration. As a main gun engagement simulator it appears quite versatile. Potential limitations are associated with its use at night where low levels of illumination or infrared engagements may pose a problem. TACED could be used as a main gun simulator on 1/20-, 1/2-, or even full-scale ranges while livefiring machinegun exercises on the latter two ranges.

# APPENDIX C

## EVALUATION OF THREE SIMULATORS IN TERMS OF SELECTED MAIN GUN ENGAGEMENTS

In the three following tables (C-1 to C-3) main gun engagements are listed from each of the three source tests. In each case the same format is followed. The first column simply indicates the number of the engagement as it appears in each source test (e.g., 1, 2, ..., etc.) and whether it is fired during the day (A) or at night (B). In the next column the engagement is provided with a number (e.g., 1, 14, 60, ..., etc.) corresponding to a particular job-objective in the gunnery domain (see Appendix A, Boldovici, Boycan, Fingerman, and Wheaton, 1979, for a detailed behavioral description of each objective). As explained in the text engagements in Tables C-2, and C-3 have been provided with alternative job-objective numbers. The next 11 columns describe each job-objective or exercise in terms of engagement conditions. Finally, the last three columns in each table contain comments as to the adequacy of each candidate device with respect to the various main gun engagements. Results of the evaluation are described below briefly for the exercises in each test.

Inspection of the comments in Table C-1 suggests the following appraisal of the candidate devices for main gun exercises in the model test. The Telfare .50 caliber device used on a 1/2-scale range can be used to simulate all main gun exercises, a most laudable outcome given that these exercises represent a great variety of engagement conditions. The TACED gun camera also fares reasonably well, providing for the simulation of 11 of the 13 primary main gun job objectives. The two exercises on which it was judged as weak are fired at night with the targets illuminated by infrared radiation. This circumstance raises a more general problem potentially limiting the utility of TACED. Its ability to deal with targets that are engaged at night under relatively low levels of illumination (e.g., as when flares are used to illuminate a target to be engaged at long range), or under infrared conditions, is problematic. To the extent that TACED can cope successfully with these potential limitations it can serve as a useful simulator for testing purposes. (A problem with TACED not mentioned heretofore is its inability to simulate subsequent rounds used in order to adjust fire in response to a first round miss. Sensing of the "round" is not feasible with TACED because, when it is used alone, no "round" is fired. The consequences of this problem in tests that are based on two-round engagements [assuming a first-round miss] have been dealt with elsewhere in the report.)

The simulation of the model test provided by the Brewster M16 device used on a 1/20-scale range is decidedly inferior. Eight of the 13 exercises cannot be simulated adequately. In two of these cases the problem is a telescopic engagement

## Table C-1.

## Analysis of Main Gun Marksmanship Exercises: Model Test

| Engagement Number | Job Objective Number | Crew Member Firing | Weapon | Mode of Firing | Firing Vehicle Motion | Target Motion | Target Type | Visibility ** | Day or Night | Fire Control Instrument | Ammunition | Target Range | DEVICES: Brewster, M16 1/20 Scale Range | DEVICES: Telfare Device 1/2 Scale Range | DEVICES: TACED 1/2 Scale Range |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A1 | 1 | G | MG | BS | MH | M | TNK | VIS | D | TEL | TPDS-T | 1600m | Target must beat zeroed range | OK | OK |
| A2 | 14 | G | MG | BS | M | S | TNK | VIS | D | GPD | HEAT-TP-T | 1000m | OK | OK | OK |
| A3 | 60 | TC | MG | P | MH | M | TNK | VIS | D | RFD | TPDS-T | 1700m | Cannot range | OK | OK |
| A4 | 43 | G | MG | P | MH | M | TNK | VIS | D | TEL | HEAT-TP-T | 1700m | Cannot range | OK | OK |
| A5 | 51 | G | MG | P | MH | S | TNK | VIS | D | GPD | TPDS-T | 2000m | Cannot range | OK | OK |
| A6 | 92 | TC | MG | P | MH | S | BKR/CRW | VIS | D | RFD | HEP-TP-T | 2200m | Cannot range | OK | OK |
| A7 | 67 | G | MG | P | MH | M | TSV | VIS | D | TEL | HEP-TP-T | 1200m | Cannot range | OK | OK |
| B1 | 113 | G | MG | RCLD | S | M | TNK | F | N | TEL | HEAT-TP-T | 1900m | Target must beat zeroed range | OK | OK |
| B2 | 103 | G | MG | RCLD | S | S | TRPS | RL | N | GPI | BEE | 900m | OK | OK | Cannot handle IR |
| (B2) | (106) | G | MG | RCLD | S | M | TSV/CRW | RL | N | GPI | HEP-TP-T | 900m | *( OK | OK | Cannot ) handle IR |
| B3 | 119 | TC | MG | RCLD | S | M | TNK | F | N | RFD | TPDS-T | 1400m | OK | OK | Maybe OK |
| B4 | 28 | G | MG | BS | S | S | TNK | RL | N | GPI | HEAT-TP-T | 800m | OK | OK | Cannot handle IR |
| B5 | 81 | G | MG | P | MH | S | TRPS | F | N | GPD | BEE | 1700m | Cannot range | OK | Maybe OK |
| (B5) | (69) | G | MG | P | MH | S | BKR/CRW | F | TEL | TEL | HEP-TP-T | 1700m | *(Cannot range | OK | Maybe OK ) |
| B6 | 31 | TC | MG | BS | M | S | TNK | F | N | RFD | TPDS-T | 1300m | OK | OK | Maybe OK |

* Exercises 106 and 69 are substitutes for the preceding BEEHIVE engagements in the event that BEEHIVE cannot be fired.

** RL = Red Light; F = Flare

## Table C-2.

## Analysis of Main Gun Marksmanship Exercises : Table VIII

| Engagement Number | Job Objective Number | Crew Member Firing | Weapon | Mode of Firing | Firing Vehicle Motion | Target Motion | Target Type | Visibility * | Day or Night | Fire Control Instrument | Ammunition | Target Range | DEVICES | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | | Brewster, M16 1/20 Scale Range | Telfare Device 1/2 Scale Range | TACED 1/2 Scale Range |
| A1 | 14 | G | MG | BS | M | S | TNK | VIS | D | GPD | HEAT-TP-T | 800-1000m | OK | OK | OK |
| | or 9 | G | MG | BS | M | S | TNK | VIS | D | TEL | HEAT-TP-T | 800-1000m | Target must be at zeroed ranged | OK | OK |
| A2a | 24 | G | MG | BS | S | M | TNK | VIS | D | GPD | TPDS-T | 1200-1600m | OK | OK | OK |
| | or 19 | G | MG | BS | S | M | TNK | VIS | D | TEL | TPDS-T | 1200-1600m | Target must be at zeroed range | OK | OK |
| A2b | 24 | G | MG | BS | S | M | TNK | VIS | D | GPD | TPDS-T | 1200-1600m | OK | OK | OK |
| | or 19 | G | MG | BS | S | M | TNK | VIS | D | TEL | TPDS-T | 1200-1600m | Target must be at zeroed range | OK | OK |
| A3 | 42 | TC | MG | BS | S | S | TNK | VIS | D | RFD | HEAT-TP-T | 800-1100m | OK | OK | OK |
| A4 | 58 | G | MG | P | S | S | TNK | VIS | D | GPD | TPDS-T | 1600-2000m | Cannot range | OK | OK |
| | or 53 | G | MG | P | S | S | TNK | VIS | D | TEL | TPDS-T | 1600-2000m | Cannot range | OK | OK |
| A5 | 75 | G | MG | P | S | S | CRW | VIS | D | TEL | HEP-TP-T | 1400-1800m | Cannot range | OK | OK |
| B1 | 11 | G | MG | BS | M | S | TNK | RL | N | GPI | HEAT-TP-T | 800-1000m | OK | OK | Cannot handle IR |
| B2a | 25 | G | MG | BS | S | M | TNK | F | N | GPD | TPDS-T | 1200-1600m | OK | OK | Maybe OK |
| | or 20 | G | MG | BS | S | M | TNK | F | N | TEL | TPDS-T | 1200-1600m | Target must be at zeroed range | OK | Maybe OK |
| B2b | 25 | G | MG | BS | S | M | TNK | F | N | GPD | TPDS-T | 1200-1600m | OK | OK | Maybe OK |
| | or 20 | G | MG | BS | S | M | TNK | F | N | TEL | TPDS-T | 1200-1600m | Target must be at zeroed range | OK | Maybe OK |
| B3 | 41 | TC | MG | BS | S | S | TNK | WL | N | RFD | HEAT-TP-T | 800-1100m | OK | OK | OK |
| B4 | 57 | G | MG | P | S | S | TNK | F | N | GPD | TPDS-T | 1600-2000m | Cannot range | OK | Maybe OK |
| | or 52 | G | MG | P | S | S | TNK | F | N | TEL | TPDS-T | 1600-2000m | Cannot range | OK | Maybe OK |
| B5 | 110 | G | MG | RCLD | S | S | CRW | F | N | TEL | HEP-TP-T | 1400-1800m | Target must be at zeroed range | OK | Maybe OK |

* RL = Red Light; F = Flare

## Table C-3. Analysis of Main Gun Marksmanship Exercises: Draft Revised Table VIII

| Engagement Number | Job Objective Number | Crew Member Firing | Weapon | Mode of Firing | Firing Vehicle Motion | Target Motion | Target Type | Visibility* | Day or Night | Fire Control Instrument | Ammunition | Target Range | DEVICES: Brewster, M16 1/20 Scale Range | DEVICES: Telfare Device 1/2 Scale Range | DEVICES: TACED 1/2 Scale Range |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A1 | 14 | G | MG | BS | M | S | TNK | VIS | D | GPD | TPDS-T | 1400-1600m | OK | OK | OK |
| | or 9 | G | MG | BS | M | S | TNK | VIS | D | TEL | TPDS-T | 1400-1600m | Target must be at zeroed range | OK | OK |
| A2a | 24 | G | MG | BS | S | M | TNK | VIS | D | GPD | TPDS-T | 1200-1600m | OK | OK | OK |
| | or 19 | G | MG | BS | S | M | TNK | VIS | D | TEL | TPDS-T | 1200-1600m | Target must be at zeroed range | OK | OK |
| | or 56 | G | MG | P | S | M | TNK | VIS | D | GPD | TPDS-T | 1200-1600m | Cannot range | OK | OK |
| | or 50 | G | MG | P | S | M | TNK | VIS | D | TEL | TPDS-T | 1200-1600m | Cannot range | OK | OK |
| A2b | 24 | G | MG | BS | S | M | TNK | VIS | D | GPD | TPDS-T | 1200-1600m | OK | OK | OK |
| | or 19 | G | MG | BS | S | M | TNK | VIS | D | TEL | TPDS-T | 1200-1600m | Target must be at zeroed range | OK | OK |
| | or 56 | G | MG | P | S | M | TNK | VIS | D | GPD | TPDS-T | 1200-1600m | Cannot range | OK | OK |
| | or 50 | G | MG | P | S | M | TNK | VIS | D | TEL | TPDS-T | 1200-1600m | Cannot range | OK | OK |
| A4a | 58 | G | MG | P | S | S | TNK | VIS | D | GPD | TPDS-T | 1800-2000m | Cannot range | OK | OK |
| | or 53 | G | MG | P | S | S | TNK | VIS | D | TEL | TPDS-T | 1800-2000m | Cannot range | OK | OK |
| A4b | 58 | G | MG | P | S | S | TNK | VIS | D | GPD | TPDS-T | 1800-2000m | Cannot range | OK | OK |
| | or 53 | G | MG | P | S | S | TNK | VIS | D | TEL | TPDS-T | 1800-2000m | Cannot range | OK | OK |
| A5a | 29 | G | MG | BS | S | S | TNK | VIS | D | GPD | HEAT-TP-T | 800-1100m | OK | OK | OK |
| | or 26 | G | MG | BS | S | S | TNK | VIS | D | TEL | HEAT-TP-T | 800-1100m | Target must be at zeroed range | OK | OK |
| | or 58 | G | MG | P | S | S | TNK | VIS | D | GPD | HEAT-TP-T | 800-1100m | Cannot range | OK | OK |
| | or 53 | G | MG | P | S | S | TNK | VIS | D | TEL | HEAT-TP-T | 800-1100m | Cannot range | OK | OK |
| A5b | 29 | G | MG | BS | S | S | TNK | VIS | D | GPD | HEAT-TP-T | 800-1100m | OK | OK | OK |
| | or 26 | G | MG | BS | S | S | TNK | VIS | D | TEL | HEAT-TP-T | 800-1100m | Target must be at zeroed range | OK | OK |
| | or 58 | G | MG | P | S | S | TNK | VIS | D | GPD | HEAT-TP-T | 800-1100m | Cannot range | OK | OK |
| | or 53 | G | MG | P | S | S | TNK | VIS | D | TEL | HEAT-TP-T | 800-1100m | Cannot range | OK | OK |
| A5c | 29 | G | MG | BS | S | S | TNK | VIS | D | GPD | HEAT-TP-T | 800-1100m | OK | OK | OK |
| | or 26 | G | MG | BS | S | S | TNK | VIS | D | TEL | HEAT-TP-T | 800-1100m | Target must be at zeroed range | OK | OK |
| | or 58 | G | MG | P | S | S | TNK | VIS | D | GPD | HEAT-TP-T | 800-1100m | Cannot range | OK | OK |
| | or 53 | G | MG | P | S | S | TNK | VIS | D | TEL | HEAT-TP-T | 800-1100m | Cannot range | OK | OK |
| B7a | 117 | G | MG | RCLD | S | S | TNK | RL | N | GPI | HEAT-TP-T | 800-1000m | OK | OK | OK |
| B7b | 117 | G | MG | RCLD | S | S | TNK | RL | N | GPI | HEAT-TP-T | 800-1000m | OK | OK | Cannot handle IR |
| B9a | 30 | G | MG | BS | S | S | TNK | F | N | GPD | TPDS-T | 1200-1400m | OK | OK | Maybe OK |
| | or 27 | G | MG | BS | S | S | TNK | F | N | TEL | TPDS-T | 1200-1400m | Target must be at zeroed range | OK | Maybe OK |
| | or 57 | G | MG | P | S | S | TNK | F | N | GPD | TPDS-T | 1200-1400m | Cannot range | OK | Maybe OK |
| | or 52 | G | MG | P | S | S | TNK | F | N | TEL | TPDS-T | 1200-1400m | Cannot range | OK | Maybe OK |
| B9b | 25 | G | MG | BS | S | M | TNK | F | N | GPD | TPDS-T | 1200-1600m | OK | OK | Maybe OK |
| | or 20 | G | MG | BS | S | M | TNK | F | N | TEL | TPDS-T | 1200-1600m | Target must be at zeroed range | OK | Maybe OK |
| | or 55 | G | MG | P | S | M | TNK | F | N | GPD | TPDS-T | 1200-1600m | Cannot range | OK | Maybe OK |
| | or 49 | G | MG | P | S | M | TNK | F | N | TEL | TPDS-T | 1200-1600m | Cannot range | OK | Maybe OK |

* RL = Red Light; F = Flare

which, for reasons cited earlier, can only be conducted at the device's zeroed range. More significantly, in the other six cases ranging is not possible for reasons given earlier. This circumstance precludes the possibility of assessing crew performance on any engagement conducted in the precision mode.

Essentially comparable results are obtained for exercises in the Table VIIIs (Tables C-2 and C-3). Telfare can provide for simulation of any and all of the engagements comprising these two tables. TACED is in fact potentially even better suited to the Table VIII exercises, because in neither the current or revised table is much emphasis placed on infrared engagements. The Brewster M16 device again appears severely handicapped. The limitation on telescopic engagements is noted again and among the alternative exercises there are several such engagements. Again, however, there are serious problems with precision exercises and both Table VIIIs provide for several of these.

# APPENDIX D

## SIMULATED MODEL TEST OF TANK CREW MARKSMANSHIP

Table D-1.

Simulated Model Test of Tank Crew Marksmanship: Daylight Engagements

| | OBJECTIVE NUMBER | CREW MEMBER | WEAPON | FIRING MODE | FIRING VEHICLE MOTION | TARGET MOTION | TARGET TYPE | TARGET VISIBILITY | DAY OR NIGHT | FIRE CONTROL INSTRUMENT | AMMUNITION | TARGET RANGE* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Replication 1 | 1 | GUNNER | MG | BS | MH | M | TNK/LAV | VIS | D/N | TEL | SAB/HEAT | 160m |
| | 14 | GUNNER | MG | BS | M | S | TNK/LAV | VIS | D/N | GPD | SAB/HEAT | 1000m |
| | 60 | TC | MG | P | MH | M | TNK/LAV | VIS | D/N | RFD | SAB/HEAT | 1700m |
| | 169 | GUNNER | COAX | NP | M | S | TROOPS | VIS | D/N | INF | COAX | 300m |
| | 43 | GUNNER | MG | P | MH | M | TNK/LAV | VIS | D/N | TEL | SAB/HEAT | 1700m |
| | 51 | GUNNER | MG | P | MH | S | TNK/LAV | VIS | D/N | GPD | SAB/HEAT | 2000m |
| | 92 | TC | MG | P | MH | S | BKR/CREW | VIS | D/N | RFD | HEP | 2200m |
| | 203 | TC | COAX | NP | MH | S | TSV | VIS | D/N | RFD | COAX | 900m |
| | 67 | GUNNER | MG | P | MH | M | TSV | VIS | D/N | TEL | HEP | 1200m |
| Replication 2 | 51 | GUNNER | MG | P | MH | S | TNK/LAV | VIS | D/N | GPD | SAB/HEAT | 2000m |
| | 60 | TC | MG | P | MH | M | TNK/LAV | VIS | D/N | RFD | SAB/HEAT | 1700m |
| | 131 | GUNNER | COAX | NP | M | M | TSV/CREW | VIS | D/N | INF | COAX | 700m |
| | 1 | GUNNER | MG | BS | MH | M | TNK/LAV | VIS | D/N | TEL | SAB/HEAT | 160m |
| | 92 | TC | MG | P | MH | S | BKR/CREW | VIS | D/N | RFD | HEP | 2200m |
| | 14 | GUNNER | MG | BS | M | S | TNK/LAV | VIS | D/N | GPD | SAB/HEAT | 1000m |
| | 238 | TC | CAL.50 | NP | MH | S | AIR | VIS | D/N | TPD | CAL.50 | 2200m |
| | 67 | GUNNER | MG | P | MH | M | TSV | VIS | D/N | TEL | HEP | 1200m |
| | 43 | GUNNER | MG | P | MH | M | TNK/LAV | VIS | D/N | TEL | SAB/HEAT | 1700m |
| Replication 3 | 92 | TC | MG | P | MH | S | BKR/CREW | VIS | D/N | RFD | HEP | 2200m |
| | 144 | GUNNER | COAX | NP | M | S | TSV | VIS | D/N | INF | COAX | 500m |
| | 60 | TC | MG | P | MH | M | TNK/LAV | VIS | D/N | RFD | SAB/HEAT | 1700m |
| | 51 | GUNNER | MG | P | MH | S | TNK/LAV | VIS | D/N | GPD | SAB/HEAT | 2000m |
| | 14 | GUNNER | MG | BS | M | S | TNK/LAV | VIS | D/N | GPD | SAB/HEAT | 1000m |
| | 246 | TC | CAL.50 | NP | M | S | LAV/CREW | VIS | D/N | TPD | CAL.50 | 1500m |
| | 1 | GUNNER | MG | BS | MH | M | TNK/LAV | VIS | D/N | TEL | SAB/HEAT | 1600m |
| | 67 | GUNNER | MG | P | MH | M | TSV | VIS | D/N | TEL | HEP | 1200m |
| | 43 | GUNNER | MG | P | MH | M | TNK/LAV | VIS | D/N | TEL | SAB/HEAT | 1700m |

* Range is ½ scale for Main Gun engagements simulated with TELFARE. Machinegun engagements are fired at scaled or actual range.

Table D-2.
Simulated Model Test of Tank Crew Marksmanship: Night Engagements

| | OBJECTIVE NUMBER | CREW MEMBER | WEAPON | FIRING MODE | FIRING VEHICLE MOTION | TARGET MOTION | TARGET TYPE | TARGET VISIBILITY | DAY OR NIGHT | FIRE CONTROL INSTRUMENT | AMMUNITION | TARGET RANGE* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 113 | GUNNER | MG | RCLD | S | M | TNK/LAV | VAL | N | TEL | SAB/HEAT | 1900m |
| | 169 | GUNNER | COAX | RCLD | S | M | TSV/CREW | VAL | N | GPI | COAX | 300m |
| | 106 | GUNNER | MG | RCLD | S | M | TSV/CREW | VAL | N | GPI | HEP | 900m |
| | 119 | TC | MG | RCLD | S | M | TNK/LAV | VAL | N | RFD | SAB/HEAT | 1400m |
| Replication 1 | 217 | TC | COAX | RCLD | S | S | TSV | VAL | N | RFI | COAX | 500m |
| | 28 | GUNNER | MG | BS | S | S | TNK/LAV | VAL | N | GPI | SAB/HEAT | 800m |
| | 69 | GUNNER | MG | P | MH | S | BKR/CREW | VAL | N | TEL | HEP | 1700m |
| | 133 | GUNNER | COAX | NP | MH | S | TSV | VAL | N | INF | COAX | 900m |
| | 31 | TC | MG | BS | M | S | TNK/LAV | VAL | N | RFD | SAB/HEAT | 1300m |
| | 119 | TC | MG | RCLD | S | M | TNK/LAV | VAL | N | RFD | SAB/HEAT | 1400m |
| | 223 | TC | CAL.50 | NP | MH | S | AIR | VAL | N | TPI | CAL.50 | 900m |
| | 113 | GUNNER | MG | RCLD | S | M | TNK/LAV | VAL | N | TEL | SAB/HEAT | 1900m |
| | 31 | TC | MG | BS | M | S | TNK/LAV | VAL | N | RFD | SAB/HEAT | 1300m |
| Replication 2 | 166 | GUNNER | COAX | NP | M | S | TROOPS | VAL | N | GPI | COAX | 700m |
| | 28 | GUNNER | MG | BS | S | S | TNK/LAV | VAL | N | GPI | SAB/HEAT | 800m |
| | 106 | GUNNER | MG | RCLD | S | M | TSV/CREW | VAL | N | GPI | HEP | 900m |
| | 193 | TC | COAX | NP | MH | M | TSV/CREW | VAL | N | RFI | COAX | 300m |
| | 69 | GUNNER | MG | P | MH | S | BKR/CREW | VAL | N | TEL | HEP | 1700m |
| | 28 | GUNNER | MG | BS | S | S | TNK/LAV | VAL | N | GPI | SAB/HEAT | 800m |
| | 124 | GUNNER | COAX | NP | MH | M | TSV/CREW | VAL | N | GPI | COAX | 500m |
| | 31 | TC | MG | BS | M | S | TNK/LAV | VAL | N | RFD | SAB/HEAT | 1300m |
| | 119 | TC | MG | RCLD | S | M | TNK/LAV | VAL | N | RFD | SAB/HEAT | 1400m |
| Replication 3 | 231 | TC | CAL.50 | NP | MH | M | LAV | VAL | N | TPI | CAL.50 | 900m |
| | 69 | GUNNER | MG | P | MH | S | BKR/CREW | VAL | N | TEL | HEP | 1700m |
| | 113 | GUNNER | MG | RCLD | S | M | TNK/LAV | VAL | N | TEL | SAB/HEAT | 1900m |
| | 243 | TC | CAL.50 | NP | S | S | AIR | VAL | N | TPD | CAL.50 | 2000m |
| | 106 | GUNNER | MG | RCLD | S | M | TSV/CREW | VAL | N | GPI | HEP | 900m |

* Range is ½ scale for Main Gun engagements simulated with TELFARE. Machinegun engagements are fired at scaled or actual range.

**Table D-3.**
**Crew Qualification Decisions**

| Number of Machinegun Exercises Performed to Standard | Number of Main Gun Exercises Performed to Standard: 0-27 | 28-35 | 36-39 |
|---|---|---|---|
| 0-11 | Unqualified | Unqualified | Unqualified |
| 12-13 | Unqualified | Marginally Qualified | Marginally Qualified |
| 14-15 | Unqualified | Marginally Qualified | Qualified |

## APPENDIX E

## RELIABILITY AND VALIDITY: IMPLICATIONS FOR EVALUATING THE SIMULATED TEST

Two concepts must be considered in empirically assessing whether or not the simulated test will serve its intended purpose: reliability and validity. These concepts can be appreciated in terms of the central theme of this report, that it would be desirable to substitute a simulated set of exercises for livefire exercises in order to determine crew marksmanship qualification. Thus, one might be able to infer whether or not a particular crew is livefire qualified from performance on a simulated test.

RELIABILITY

Any measurement carried out in the real world has some chance of error. Most of us experience examples every day. To find out what we weigh for instance, we might step onto a scale. However, unless our scale is exceptional, the weight that we get when we step on it once will not be the weight that we get when we step on it a second time. To get a reasonable estimate of our weight, therefore, we might take the average of the weights that we obtain in five weighings. We have made an assumption in doing so: while any single weighing has a chance of being in error, these errors are random and will average out over several "trials." Some rather subtle concepts are involved in even this simple example. In psychometric language one speaks of a "true" weight as distinct from a "measured" weight. To the extent that the scale measures the true weight in an unbiased fashion, the error of measurement is random and an average of several measured weights is a good estimate of the true weight. However, some scales might have more error than others, i.e. they may vary more from trial to trial. The less error of measurement that a scale has, the more reliable it is considered to be.

"In brief, the reliability of a test is a measure of the consistency with which the test procedure establishes the scores" (Rozeboom, 1966, p. 375). Consistency in this context might be with regard to the true score, that is, the extent to which the measured score is consistent with the true score. Alternatively, it might be with regard to repeated measurements: to what extent does one measurement agree with a second, a third, or a fourth, using the same test or scale (since all are assumed to be estimates of the true score or weight). The notion of reliability is not only important in terms of measurement error but also, as will be discussed below, in terms of the ceiling it places on the validity of any test. A test which produces scores with large errors of measurement can never be perfectly valid. Reliability as defined above refers to consistency of measurement. If an individual receives the same score on a test, no matter how many times he takes it, then the test is reliable with regard to that individual's score. If a test is administered several times to a group of people and each individual receives the same score on each administration, then the test is reliable for that population. Immediately, one way of measuring reliability becomes apparent. One can administer the same test twice, to a single group of individuals, and compare the scores from each administration. If each individual received exactly the same score on both test trials, then the test would be perfectly reliable. This outcome, however, would be extremely unusual. More likely, individual scores would vary somewhat from administration to administration. If scores for individuals on the two administrations

were totally unrelated, then the test would have no reliability. This implies that there must be a scale of reliability that ranges from perfect to none. Such in fact is the case inasmuch as the usual statistical coefficient used to measure reliability, the reliability index, ranges from one (perfect) to zero (none), and is a form of the correlation coefficient.

The reliability index has many versions, all of which share one key characteristic: they are sensitive to the extent to which the two administrations of the test order the testees consistently. That is, they measure the extent to which testees who score well on the first administration also score well on the second, and the extent to which testees who score poorly on the first administration also score poorly on the second. The correlation coefficient by itself is not sensitive to absolute differences in scores. Thus, if all of the testees were to improve by some fixed amount on the second administration, this would have no impact on a correlational measure of reliability. This characteristic is often quite desirable. It is reasonable to think of a test not only as a test but also as a learning experience: thus, testees might improve their performance on the second administration as a result of having attempted the test items previously. Particularly in the area of psychomotor performance, as in tank gunnery, an administration of the test serves not only to assess performance but also gives the crew members a chance to practice their skills; the gunner to track, the driver to drive, and so on. Nevertheless, it would be hoped that a second administration of the test would only change performance by a fixed and equal amount for each crew being tested. Then the crews would still be ordered in the same way, even though their absolute performance level might have improved from the first to the second administration.

Suppose, however, that some crews perform very well on the first administration, passing approximately 90% of the main gun engagements (11 or 12 of 13). These crews would have very little room for further improvement. Other crews might perform less well on the first administration, perhaps passing only 50% (six or seven) of the 13 engagements. These crews have more room for improvement, and may benefit more from the training provided by the first administration. Thus, only a few of the crews who perform well on the first administration may improve measurably on the second, since improvement to perfect performance is extremely difficult, while many crews who perform at a lower level on the first administration may do appreciably better on the second administration. In this kind of situation the crews would be clearly ordered by the first administration of the test, but might be randomly bunched around the 90% level of performance on the

second. The result would be an unfairly low estimate of test reliability. Such an effect, produced by the artifactual ceiling on performance (i.e. 100%), is referred to as a "learning ceiling" effect and illustrates that a test-retest mode of assessing reliability may not always be appropriate.

In order to overcome the learning ceiling effects that may be associated with the test-retest method of assessing reliability, another method, termed the "alternate parallel forms" approach, is often employed. With this method two versions of the test are created which are assumed to be equivalent, although the content of the two versions is not identical. Testees then receive the two forms of the test (perhaps half in the order "AB" and half in the opposite order "BA") and the resulting scores are compared using the correlation coefficient. Once again, the question at issue is one of consistency. In this case, do the two alternate forms order the crews consistently? One would hope that because the content of the alternate forms was not identical, item-specific learning and consequent ceiling effects would be less severe, and better estimates of the true reliability of the test would be possible. Unfortunately, it is often extremely difficult to achieve the goal of exactly equivalent (but non-identical) parallel forms of a test. Further, in the case of tank gunnery where the items of the test are gunnery engagements, even though different engagements might be used in the two forms, the underlying behavioral skills utilized by the crew would be nearly the same from form to form. The gunner would still track with the main gun, the driver would still drive, and so on. Thus, in practice, such an approach does not necessarily eliminate the problem of learning effects.

A third method is often used to assess test reliability. It represents a compromise between the test-retest and the alternate parallel forms procedures. This approach is generally known as the "split-half" method of assessing reliability. To illustrate it, suppose that a 20-item test were constructed and the items were randomly ordered for presentation to the testee. After the testees had taken a single administration of this 20-item test, a data analyst would proceed to analyze the scores as if two parallel 10-item tests had been taken. He might, for example, use performance on the odd-numbered items (first, third, fifth, etc.) to determine a score on form A of the test, and performance on even-numbered items (second, fourth, etc.) to compute a score on form B. Note that by taking alternating items from the test sequence, he attempts to control for learning effects: equal numbers of items on forms A and B have been taken from early and late portions of the test, and thus have been preceded by approximately the same amount of practice. Scores from form A and form B could then be correlated to assess the test's reliability. There are even procedures for predicting the reliability of a full

20-item test, using the split-half correlation where each half contained 10 items.[1]

The reliability attributed to a test by this means would depend to some extent, of course, on the particular half of the items that were split into form A and form B. Various procedures have been employed, including using all of the possible split-halves and selecting the largest reliability coefficient or computing the reliability coefficients from all the possible split-halves and averaging them. Other approaches have also been considered; for example, splitting the test into three or more components and computing several reliability coefficients among each of the components and then averaging these.

Each of these basic approaches to determining test reliability deals with the internal consistency of the test, that is, the extent to which each item in the test consistently measures the same aspect of performance that all the other items in the test measure. In the case of tank gunnery marksmanship, to the extent that all exercises included in a gunnery test measure the underlying skill "marksmanship," the test should be internally consistent.[2]

While at present there are several measures of internal consistency in use, the one most widely used is the "alpha coefficient" (Cronbach, 1951). This coefficient does not require the selection of any particular split-half or group of split-halves since it is a summary measure. Since the alpha coefficient depends on both the internal consistency of the test and test length, it is not a pure measure of

---

[1] A 20-item test is generally more reliable than a 10-item test, just as averaging 20 scale weighings is likely to provide a more accurate estimate of true weight than 10 weighings. The simple split-half correlation is thus something of an underestimate for the full 20-item test. To deal with this problem the Spearman-Brown formula is used to correct for test length (Rozeboom, 1966, p. 404).

[2] The model test is not completely homogeneous in content. For example, earlier analyses (Wheaton et al., 1978) revealed that main gun and machinegun exercises involved different behavioral components. Similarly, performance on exercises where the gunner fires will not necessarily be consistent with those where the tank commander fires. Thus, internal consistency as a measure of reliability might only be appropriate within various components of the test (e.g., gunner firing main gun, or tank commander firing .50 caliber).

reliability. But often it may be used in concert with other techniques to obtain an accurate estimate of a test's true reliability. Since the alpha coefficient is based on a single sampling of the item error variance, it will generally overestimate the true reliability. When only one sample is drawn (as is the case in the nonreplicated experimental design discussed later), the alpha coefficient serves as the best available estimate. When two (or more) samples are available, both the alpha coefficient and other methods such as test-retest may be used to bound the estimate of reliability (c.f. Rozeboom, pp. 411-415, 427-496; Cronbach, Gleser, Nanda, & Rajaratnam, 1972).

In addition to methods for assessing reliability, it is also important to consider factors associated with actual test administration that may enhance or reduce reliability. Recognizing and understanding such factors is important if one is to gather test scores which have the highest possible reliability. Any aspect of test administration which prevents the testee from demonstrating his true proficiency contributes to inconsistency, thereby reducing reliability (and as pointed out above placing a ceiling on test validity). For example, administering a test with many irrelevant distractions in the test environment can contribute to inconsistency. For this reason, any test conducted in a real-world environment (e.g. gunnery marksmanship) is likely to be less reliable than a test conducted under carefully controlled conditions (e.g. a classroom test of intelligence). There is no perfect solution to this problem, since distracting features such as noise and concussion are part of the tank environment. What is clear, however, is that the control of other irrelevant and inappropriate distractions is all the more important. Tank crew fatigue may be one such distraction. Crew member illness is another. Firing under degraded weather conditions might be a third. This does not mean that one should not be interested in the performance of fatigued or ill crews, or in the effects of reduced visibility. Rather, measurements of the crew's performance for marksmanship qualification should not be obtained under these extraneous and irrelevant conditions.

Measurement error introduced by scoring procedures is another major source of unreliability in the tank gunnery test environment. The simplest example is in the scoring of target hits. Suppose that observers using BC scopes are to determine whether or not a hit is achieved. Distractions, fatigue, or illness among the observers could make their reports inconsistent, and this would certainly be an unwanted source of unreliability. Therefore, to the extent that scoring can be relegated to automatic systems which are less prone to the vagaries of human performance, measurement error

can be reduced. Wheaton et al. (1978) have discussed the importance of such controls in marksmanship testing, and this topic is returned to in a later section on implementing the empirical evaluation.

This discussion of reliability has indicated its central role in an empirical evaluation of a simulated test. Several methods for measuring reliability have been introduced, and a concern has been raised for improving test reliability and decreasing error of measurement by controlling the conditions of testing. Empirical measurement of reliability is addressed elsewhere in the report in the context of specific evaluation designs; control of test conditions is discussed again as part of the guidelines for conducting the evaluation study.

VALIDITY

Generally, one is not interested in the specific number that is obtained when a test is scored, but rather in the inferences that can be drawn from that number. Such inferences may be correct or incorrect, a situation giving rise to the validity issue. If a work-sample test is properly constructed and is highly reliable, then one may validly draw inferences from test scores as to the true performance capabilities of the testee. The statement "draw inferences from ... to ..." makes explicit the concept that one infers from a test score to a criterion, an error-free or true index of performance, one for which the test is a surrogate. Such an ultimate criterion is rarely obtainable in the real world. However, under certain circumstances good empirical approximations of such criteria are available. In the present case, for example, the true criterion is marksmanship performance on the domain of tank gunnery objectives. Since the model set of tank gunnery exercises was developed as a job-relevant work-sample test (Guion, 1978c; Guion & Ironson, 1978) it can be assumed that performance on the model set of livefire exercises is a reasonable approximation of the underlying and more general marksmanship skill associated with the domain.

There are rational approaches to validity assessment which are very important in the present context. Just as the validity of the livefire exercises was provided for by consideration of content and job relevance, the content and job relevance of the exercises in the simulated table may be examined. These rational methods of establishing validity are often as important as the empirical ones in the context of criterion-referenced testing, and set the stage for empirical assessments of validity as necessary but primarily confirmatory exercises.

At issue is whether one can reach a qualification decision from simulated test scores that is consistent with the qualification decision that would have been made from livefire test scores. In this sense it is both reasonable and necessary to demand that a crew which "qualifies" on one test should also "qualify" on the other. If both tests were administered to the same crew, four possible outcomes could be obtained. The crew could be considered qualified on both, unqualified on both, qualified on the simulated test but not on the livefire test, or vice versa. Only the first two of these outcomes are desirable since the two tests must provide equivalent and consistent information. The degree of consistency between the outcomes on the simulated and livefire tests can be defined as the validity of the simulated test.

As defined above, the concept of test validity raises two key issues. First, if the content of the simulated test is essentially the same as the content of the livefire test, why would they not lead to consistent scores? The answer is that the test content may not be identical in important ways; the livefire "flash/bang" effect, for example, may be particularly important for new gunners, but presumably would be absent from the simulated test. Suppose that a new gunner were tested on the simulated test and then tested on the livefire test. Upon his first exposure to the loud noise and recoil of main gun firing his performance on the livefire test might suffer, even if his performance in the preceding simulated test had been outstanding. The second issue, therefore, is how to establish the validity of the simulated test if one cannot simply assume it. The answer is complicated since there are many meanings of the concept of validity. Similarly, while empirical methods of establishing validity are available, selection of the correct experimental design is not always simple.

Generally, if one has an empirical measure of true performance, it is possible to describe the validity of a simulated set of exercises. The relationship between scores obtained on the simulated and livefire sets of exercises is representative of the validity. Once again, the concept of correlation comes into play. The validity of the simulated test may be characterized by correlating crew performance on the two sets of exercises. However, some of the same problems that were raised in the context of test-retest methods for determining reliability arise in estimating validity. Learning effects, in particular when they are associated with ceiling effects, may provide underestimates of validity. One method of dealing with the potential influence of such effects would be to alternate the order of testing. Thus, some of the testees (usually half) would receive the livefire criterion test first, followed by the simulated test; other crews would receive the two tests in reverse order.

## RELIABILITY AND VALIDITY IN CRITERION-REFERENCED TESTING

Traditional methods of assessing reliability and validity have a long history and much of the preceding discussion was layed out well before the advent of criterion-referenced testing (see Rozeboom for a review of traditional approaches). Criterion-referenced testing, however, introduces certain complications that must be considered in the present context. These complexities are easiest to see with regard to traditional methods for assessing the validity of tests. As pointed out above, a common procedure for assessing validity is to estimate the correlation between the test and criterion scores. The correlation, however, is sensitive only to the consistency with which the two sets of scores order testees. It is not sensitive to whether or not testees receive exactly the same score. In criterion-referenced testing the precise score is of much interest. For example, the livefire portion of the model set of gunnery exercises requires that a crew pass at least 12 of 13 main gun exercises in order to be judged qualified. If a validity study of the simulated table yielded a correlation coefficient of 1.0 (perfect correlation) it would not necessarily mean that a score on the simulated test corresponded precisely to one on the livefire criterion test. For example, performance could be relatively lower on the simulated test for all crews. As mentioned above, a constant difference between two sets of scores does not affect the correlation coefficient at all. Therefore, to the extent that the objective is literally to substitute the simulated set of exercises for the livefire set of exercises, establishing the traditional (i.e. correlational) validity of the test is not sufficient. An additional requirement is that the simulated set of exercises and their scores be calibrated to the livefire criterion test. In other words, a translation technique must be developed to determine the precise score that would have been obtained on the livefire set of exercises from the score which was obtained on the simulated set of exercises. If the two sets of scores differ by a constant, for example, then it would be important to determine this constant. Other translations are also possible, as, for example, when performance on a single simulated exercise is perfectly predictive of the qualification decision made on the livefire set of exercises. A great deal of sophisticated mathematical treatment has been applied to this calibration problem (Rasch modeling, latent trait analysis, scaling procedures, etc.). The reader is referred to a recent series of reports by Guion (1978a, b, c; Guion & Ironson, 1978) for a review. The important point is that empirical data must be acquired in order to establish the calibration, making it possible to translate from performance on the simulated version of the test to performance on the livefire version.

Criterion-referenced testing also introduces some wrinkles into the assessment of reliability. For example, in using the correlation coefficient to determine reliability, one needs variance among the scores. That is, some testees must do well on the test and some testees must do poorly. If they all obtain the same score, so that there is no variance, then they are not ordered on the test and the correlation coefficient (only sensitive to the consistency with which their scores are ordered) is zero. Variances generally are no problem in traditional testing (e.g., I.Q. or school achievement). However, in criterion-referenced testing of tank gunnery marksmanship it may be a problem. Highly trained crews are not likely to vary to any great extent in their performance on the simulated or livefire test. If their training has been successful, it is likely that their scores will cluster around (hopefully above) the cutoff that has been established for qualification. Fortunately, the problem of invariance can be dealt with in a number of ways. There are, for example, special mathematical treatments for determining the reliability of criterion-referenced tests which may be of help (Livingston, 1972a, b, c; Harris, 1972; Brennan & Kane, 1977; Cronbach, Gleser, Nanda, & Rajaratnam, 1972).

Assessment of the reliability of criterion-referenced tests may also be improved by using a finer level of measurement. Instead of simply scoring the test or each item on a go/no-go basis, one may acquire data on the underlying performance measures which define the go/no-go decision. In the context of tank gunnery this might include measuring the actual elapsed time to fire and/or the distance between the strike of the round and the center of the target, instead of simply noting whether or not firing was accomplished in a specified amount of time and whether or not a hit was achieved (see also Wheaton, et al., 1978). With more detailed measurement, more score variance is likely.[3]

Given the need to establish empirically the reliability and validity of a simulated set of exercises and to calibrate them vis à vis the livefire marksmanship test, other portions of the report develop various methods for collecting the necessary data. Consideration is given not only to establishing these measures with psychometric purity, but also to the practical realities of such testing. Testing in a tank

[3] Reliability of these more finely detailed measures would be sufficient, but not necessary for reliability of item go/no-go or whole-test qualification scores. Most test developers would agree, however, that reliability of finer measures would be a useful, albeit stringent, way to look at test reliability.

environment is expensive. Thus, it is important to design one or more evaluation experiments that will provide for an adequate assessment of reliability and validity while at the same time consuming no more than a reasonable amount of resources.